STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
January 13, 2021

Categorical Data Analysis – Home Examination

January 13, 2021, 9.00-16.00

Examination by: Ola Hössjer, ph. 070 672 12 18, ola@math.su.se
Allowed to use: Miniräknare/pocket calculator (including the use of R as pocket calculator), table in the appendix of this exam, course literature and other course material. You are not allowed to ask anyone for help.
Inlämning/Handing in: The solutions should by sent by email as a pdf file to the eaminator by 16.00. Either scanned hand-written notes, or a pdf file generated from a word processor. *Återlämning/Return of exam:* Will be communicated by email.

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 at first. Exercises need not to be ordered from simpler to harder.

Problem 0

Verify that you have solved all exerices without help from anyone. This is required in order for the solutions to be corrected. (0p)

Problem 1

Let Y equal 1 or 0 depending on whether a person of age x years has ever had any coronary heart disease (CHD) symptoms or not. Hosmer and Lemeshow (1989) describe a study with 100 subjects that reported their age x_i and CHD status y_i, $i = 1, \ldots, 100$.

a. Formulate a logistic regression model for the probability $\pi(x; \alpha, \beta) = P(Y = 1|x)$ that an x year old individual has experienced CHD symptoms. (2p)

b. Parameter estimates where $\hat{\alpha} = -5.310$ (intercept) and $\hat{\beta} = 0.111$ (effect parameter). Use this to in order to estimate the odds ratio of having experienced CHD symptoms, between two individuals who are 65 and 45 years of age. (2p).

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 in order to compute a 95% confidence interval first for the logarithm of the odds ratio in 1b), and then a confidence interval for the odds ratio itself. (2p)

d. When the participants of the study are clumped into 4 age groups $a = 1, 2, 3, 4$, the n_a members of each age group are assumed to have the same age x_a, with n_{a0} and n_{a1} the number of individuals in group a having $Y = 0$ and $Y = 1$ respectively, and with $\hat{\pi}_a = \pi(x_a; \hat{\alpha}, \hat{\beta})$ the logistic regression model fit. This is reported in the following table:

Group a	x_a	n_a	n_{a0}	n_{a1}	$\hat{\pi}_a$	$n_a\hat{\pi}_a$
20-34	28	25	22	3	0.100	2.49
35-44	40	27	19	8	0.295	7.97
45-54	50	21	10	11	0.560	11.75
55-64	60	27	6	21	0.794	21.44
Sum		100	57	43		

Use data from this table in order to perform a X^2 goodness-of-fit test of the logistic regression model, at significance level 5%. (4p)

Problem 2

Consider two categorical variables X and Y with levels $1 \leq i \leq I$ and $1 \leq j \leq J$. These two variables are registered for n individuals according to multinomial sampling. The result is reported in a twoway $I \times J$ contingency table, with cell counts N_{ij}, for $1 \leq i \leq I$ and $1 \leq j \leq J$.

a. Write down the log likelihood function in terms of the joint cell probabilities $\pi_{ij} = P(X = i, Y = j)$ and observed cell counts n_{ij}. How many free parameters are there? (2p)

b. Formulate the null hypothesis H_0 of independence between X and Y, against the alternative hypothesis H_a of non-independence. Phrase these hypotheses in terms of the joint cell probabilities π_{ij}, and the marginal probabilities π_{i+} and π_{+j}. (2p)

c. The proportional reduction

$$U = \frac{\sum_{i,j} \pi_{ij} \log(\pi_{ij}/(\pi_{i+}\pi_{+j}))}{-\sum_{j=1}^{J} \pi_{+j} \log(\pi_{+j})}$$

in entropy is a measure of dependency between X and Y, which tells how much our knowledge of X reduces the uncertainty (entropy) of Y. It ranges between $U = 0$ (independence between X and Y) to $U = 1$ (full deterministic dependency $Y = f(X)$). Let $\hat{U}$ be an estimate of U, defined by replacing cell probabilities with estimates $\hat{\pi}_{ij} = n_{ij}/n$, $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$ respectively. Motivate why approximately

$$-2n\hat{U}\sum_{j=1}^{J}\hat{\pi}_{+j}\log(\hat{\pi}_{+j}) \overset{H_0}{\sim} \chi^2_d \tag{1}$$

for large datasets n and some number d of degrees of freedom. In particular, find this number d. (3p)

d. In a US survey from 2006, $n = 1009$ individuals of different age were asked how much they liked their jobs. The result of the study is summarized in the following table:

	Job Satisfaction Y		
Age X	$j = 1$ (=not satisf)	$j = 2$ (=fairly satisf)	$j = 3$ (=very satisf)
$i = 1$ (< 30	34	53	88
$i = 2$ (30-50)	80	174	304
$i = 3$ (> 50)	29	75	172
$\hat{\pi}_{+j}$	0.142	0.299	0.559

It was found that $\hat{U} = 0.0052$. Use this, data from the table and Problem 2c) in order to test, at significance level 5%, whether job satisfaction is independent of age. (3p)

Problem 3

An investigation of mortality in leukemia was conducted among survivors of the atom bomb 1945 in Hiroshima. Individuals were categorized according to their age group Z, their radiation dose X and whether they died in leukemia or not (Y) within a certain number of years. This is summarized in the threeway contingency table below. It is assumed that the cell counts of this table are observations of independent Poisson distributed random variables.

	Did not die in leukemia ($j = 1$)		Died in leukemia ($j = 2$)	
Age	Low dose ($i = 1$)	High dose ($i = 2$)	Low dose ($i = 1$)	High dose ($i = 2$)
0-20 years ($k = 1$)	39 160	3 882	25	26
20-50 years ($k = 2$)	41 664	4 291	39	26
50- years ($k = 3$)	15 163	1 337	13	10
Sum	95 987	9 510	77	62

a. The table below displays the deviance $G^2(M)$ of four loglinear and nested models M. Select the best model, among these four models, using Forward Inclusion, where models are tested pairwise at significance level 5%. (Hint: Start by computing the number of parameters of each model.) (4p)

M	$G^2(M)$
(XY, XZ, YZ)	1.67
(XY, XZ)	2.69
(XY, Z)	25.42
(X, Y, Z)	147.00

b. Let n_{ijk} and μ_{ijk} be the observed and expected count of a cell with $X = i$, $Y = j$ and $Z = k$, so that, for instance, $n_{122} = 39$ is the number of individuals of age 20-50 years with a low radiation dose who died of leukemia. Find an expression (no proof is required) for the expected cell count μ_{122} for each one of the models $M_1 = (XY, Z)$ and $M_2 = (XY, XZ)$, in tems of appropriate marginals μ_{ij+}, μ_{i+k}, μ_{i++}, μ_{++k}, and μ_{+++} of the expected cell counts. Then find the maximum likelihood estimate $\hat{\mu}_{122}$ of μ_{122}, for each one of M_1 and M_2. You may use that $n_{+++} = 105636$. (4p)

c. Compute, for model M_1, a maximum likelihood estimate $\hat{\theta}^{XY}_{(k)}$ of the conditional odds ratio

$$\theta^{XY}_{(k)} = \frac{P(Y = 2|X = 2, Z = k)/P(Y = 1|X = 2, Z = k)}{P(Y = 2|X = 1, Z = k)/P(Y = 1|X = 1, Z = k)}$$

of dying in leukemia between individuals with a high and low radiation dose, given that they belong to age group k. (2p)

Problem 4

We continue studying the dataset of Problem 3. As in 3c), we are primarily interested in the effect that radiation dose has on leukemia mortality. Thus we treat death Y in leukemia as an outcome variable, radiation dose X as a predictor and age Z as a confounder.

a. Define the loglinear parameters of $M_2 = (XY, XZ)$ and $M_3 = (XY, XZ, YZ)$ respectively. In particular, specify for each model which parameters you put to zero in order to avoid overparametrization. (3p)

b. Show, for model M_3, that $P(Y = 2|X = i, Z = k)$, the conditional probability of death in leukemia given radiation dose and age, defines a logistic regression model. Express its parameters as functions of the loglinear parameters from 4a). (3p)

c. It can be shown that M_2 also gives rise to a logistic model for $P(Y = 2|X = i, Z = k)$. Which of the two logistic regression models, derived from M_2 and M_3 respectively, is selected by Akaike's Information Criterion AIC? (Hint: Data from Problem 3 will be helpful, but a full score requires consideration of the log likelihood of a logistic regression model.) (4p)

Problem 5

Return to the age-grouped CHD dataset of Problem 1d). The Cochran-Armitage (CA) trend test will be used to test the two hypotheses

$$\begin{aligned} H_0: &\quad \beta = 0, \\ H_a: &\quad \beta > 0, \end{aligned}$$

whether age increases the risk of having experienced CHD symptoms or not. Although it is intuitively clear that H_a should hold, we will check if the dataset is large enough to warrant such a conclusion, at significance level 0.1 %. The CA test is based on the score statistic

$$z_S = \frac{\sum_{a=1}^{4}(x_a - \bar{x})n_{a1}}{\sqrt{p(1-p)\sum_{a=1}^{4} n_a(x_a - \bar{x})^2}}, \tag{2}$$

where

$$\begin{aligned} \bar{x} &= \textstyle\sum_a n_a x_a / n = 44.5, \\ p &= n_{+1}/n = 43/100 = 0.43, \end{aligned}$$

is the average age and average fraction of individuals that ever experienced CHD sympoms, respectively. The two sums of the test statistic equal

$$\begin{aligned} \textstyle\sum_{a=1}^{4}(x_a - \bar{x})n_{a1} &= 300.5, \\ \textstyle\sum_{a=1}^{4} n_a(x_a - \bar{x})^2 &= 14475. \end{aligned}$$

a. Perform the CA test at level 0.1 %. (Hint: You may use the standard normal quantile $z_{0.001} = 3.09$.) (2p)

b. Define the log likelihood $L(\alpha, \beta)$ of the age group data set in terms of n_a, n_{a1}, and $\pi_a = \pi(x_a; \alpha, \beta)$ for $a = 1, \ldots, 4$. (2p)

c. Use 5b) in order to find expressions for the two components

$$\begin{aligned} u_\alpha(\alpha, \beta) &= \partial L(\alpha, \beta)/\partial\alpha, \\ u_\beta(\alpha, \beta) &= \partial L(\alpha, \beta)/\partial\beta \end{aligned}$$

of the score function. (2p)

d. Use 5c) to derive the two diagonal elements $J_{\alpha\alpha}(\alpha, \beta)$ and $J_{\beta\beta}(\alpha, \beta)$ of Fisher's information matrix, as well as the value $J_{\alpha\beta}(\alpha, \beta)$ of the two non-diagonal elements of this matrix. (2p)

e. Verify that (2) is indeed a score statistic, by showing that

$$z_S = \frac{u_\beta(\hat{\alpha}(0), 0)}{\sqrt{\mathrm{Var}[u_\beta(\hat{\alpha}(0), 0)]}},$$

where $\hat{\alpha}(\beta)$ is the value of α that maximizes the log likelihood when β is fixed. (Hint: You may use that $\pi(x_a; \hat{\alpha}(0), 0) = p$ for all age groups a, and the formula

$$\mathrm{Var}[u_\beta(\hat{\alpha}(\beta), \beta)] = J_{\beta\beta}(\hat{\alpha}(\beta), \beta) - \frac{J_{\alpha\beta}(\hat{\alpha}(\beta), \beta)^2}{J_{\alpha\alpha}(\hat{\alpha}(\beta), \beta)}$$

for the variance of the score function.) (2p)

Good luck !

Appendix A - Table for chi-square distribution

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