STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
February 23 2023

Categorical Data Analysis – Examination

February 23, 2023, 8.00-13.00

Examination by: Ola Hössjer, ph. 070 672 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 through the course page.

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

A new medicine against severe hypertension (högt blodtryck) was tested in a study where 200 patients were given daily dosages $x \geq 0$ mg for one year, with $x = 0$ corresponding to placebo. The study was blind, so that no patient knew of their dosage x. It was registered whether they suffered from a heart attack during the study period ($Y = 1$) or not ($Y = 0$). The estimates of the intercept and slope parameters of a logistic regression model where $\hat{\alpha} = -3.1$ and $\hat{\beta} = -0.32$. They are jointly approximately normally distributed, with an estimated covariance matrix

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) \\ \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) & \widehat{\mathrm{Var}}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} 1.1 & -0.06 \\ -0.06 & 0.0225 \end{pmatrix}.$$

a. Write down $\pi(x) = P(Y = 1|x)$ for a logistic regression model. (1p)

b. Perform a *one-sided* test to check if the medicine has any *preventive* effect. (Hint: The 0.95 quantile of a standard normal is $z_{0.05} = 1.645$.) (3p)

c. What is the predicted probability $\pi(10)$ of heart attack with a daily dosage of 10 mg? (2p).

d. Compute a 95% confidence interval for $\pi(10)$. (Hint: Start by constructing a confidence interval for $\text{logit}[\pi(10)]$.) (4p).

Problem 2

For a cross-sectional study with Poisson sampling, the severity of injury and the health status one year later was investigated for 35 drivers involved in car accidents. The observed values n_{ij} of the number of drivers $N_{ij} \sim \text{Po}(\mu_{ij})$ are summarized for all rows i and columns j of the following table:

	Health status j			
Injury i	Good (=1)	Fair (=2)	Bad (=3)	Total
None (=1)	7	4	1	12
Mild (=2)	5	7	2	14
Fatal (=3)	1	2	6	9
Total	13	13	9	35

a. Compute the odds ratio estimates for the following 2×2 subtables:

$$\begin{aligned} \text{I:} &\quad \{\text{None}, \text{Mild}\} \times \{\text{Good}, \text{Fair}\}, \\ \text{II:} &\quad \{\text{None}, \text{Fatal}\} \times \{\text{Good}, \text{Bad}\}, \\ \text{III:} &\quad \{\text{Mild}, \text{Fatal}\} \times \{\text{Fair}, \text{Bad}\}. \end{aligned}$$

What are your conclusions? (2p)

b. The investigator wanted to look more closely into the impact a mild injury had on health, and decided to discard the third row and column, and only analyze the upper 2×2 table I. Formulate, in terms of an odds ratio, the null hypothesis H_0 that a mild accident has no effect on health status one year later for this subtable. (2p)

c. Fisher's exact test of the null hypothesis from 2b uses only N_{11}, and is based on a certain conditional distribution $P_{H_0}(N_{11} = n_{11}| \ldots)$, displayed below. Determine the condition (the dots) and write down the formula for this conditional distribution (you don't have to prove it). (3p)

n_{11}	0	1	2	3	4	5
$P_{H_0}(N_{11} = n_{11}\| \ldots)$	0.0000	0.0001	0.0027	0.0268	0.1208	0.2706
n_{11}	6	7	8	9	10	11
$P_{H_0}(N_{11} = n_{11}\| \ldots)$	0.3157	0.1933	0.0604	0.0089	0.0005	0.0000

d. Suppose we want to test H_0 in 2b against a one-sided alternative hypothesis H_a that a mild injury increases risk of impaired health. Formulate H_a in terms of an odds ratio. Compute the corresponding one-sided P-value for the given data set, using Fisher's exact test. (3p)

Problem 3

Consider the 2×2 subtable of Problem 2b-d. Although this is not a clinical trial, we can still condition on the row sums and treat it as such, with $\pi_1 = \pi_{1|1} = \pi_{11}/(\pi_{11} + \pi_{12})$ and $\pi_2 = \pi_{1|2} = \pi_{21}/(\pi_{21} + \pi_{22})$ the conditional probabilities of being healthy one year after the car accident, given that the driver had no injury and a mild injury respectively. With this conditioning we thus have independent binomial rows sampling.

a. Write down the likelihood of data in terms of π_1 and π_2. (2p)

b. Define the relative risk r and write down the hypotheses of a two-sided test where the null hypothesis asserts that a mild injury has no effect on health status one year later. (2p)

c. Use 3a to prove that the likelihood ratio test statistic for the hypotheses in 3b is

$$G^2 = 2\left(n_{11}\log\frac{n_{11}/n_{1+}}{n_{+1}/n} + n_{12}\log\frac{n_{12}/n_{1+}}{n_{+2}/n} + n_{21}\log\frac{n_{21}/n_{2+}}{n_{+1}/n} + n_{22}\log\frac{n_{22}/n_{2+}}{n_{+2}/n}\right),$$

where $n = n_{++}$ is the total number of observations of the 2×2 table. (Hint: You may without proof use that the ML estimates of π_1 and π_2 are n_{11}/n_{1+} and n_{21}/n_{2+} under the alternative hypothesis, whereas n_{+1}/n is an ML-estimate of the probability of being healty under the null hypothesis.) (3p)

d. Use G^2 in order to perform a two-sided likelihood ratio test at level 5% in order to check if a mild injury has any effect on health status one year later. (Hint: You don't need to have verified the formula for G^2 in 2c order to solve 2d.) (3p)

Problem 4

A threeway table contains data of the binary categorical variables X, Y and Z. The number of observations $N_{ijk} \in \text{Po}(\mu_{ijk})$ with $X = i$, $Y = j$ and $Z = k$ is Poisson distributed for $1 \leq i, j, k \leq 2$, and independent for all different cells (i, j, k).

a. Let (XY, Z) be the loglinear model where Z is jointly independent of X and Y. Express all expected cell counts μ_{ijk} in terms of the loglinear parameters, excluding those that are put to zero in order to avoid overparametrization. (3p)

b. Use part 4a to prove that

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}},$$

where a plus sign denotes summation over the corresponding index. (2p)

c. Use 4b and data n_{ijk} from the two partial tables below to find the ML estimates $\hat{\mu}_{ijk}$ of all μ_{ijk}. (Hint: It will be helpful to use the observed values $n_{11+} = 72$, $n_{12+} = 119$, $n_{21+} = 32$ and $n_{22+} = 239$ of the marginal table of X and Y. The total sizes of the two partial tables are $n_{++1} = 168$ and $n_{++2} = 294$.) (2p)

Observed values n_{ij1}:

	$j=1$	$j=2$
$i=1$	25	38
$i=2$	12	93

Observed values n_{ij2}:

	$j=1$	$j=2$
$i=1$	47	81
$i=2$	20	146

d. Choose between (XY, Z) and the saturated model (XYZ) using Akaike's information criterion AIC. (3p)

Problem 5

Consider the 2×2 table of Problem 2b-d, with injuries and health status of drivers. We assume independent binomial rows, as in Problem 3, but with a logistic parametrization

$$\begin{array}{rcl} \pi_1 & = & \exp(\alpha)/[1+\exp(\alpha)], \\ \pi_2 & = & \exp(\alpha+\beta)/[1+\exp(\alpha+\beta)] \end{array} \tag{1}$$

for the probabilities π_1 (π_2) that drivers with no (a mild) injury have good health status one year later.

a. Express the log likelihood $L(\alpha, \beta)$ in terms of the parameters α, β, the cell counts $\{n_{11}, n_{12}, n_{21}, n_{22}\}$ and the row sums $n_1 = n_{11} + n_{12}$ and $n_2 = n_{21} + n_{22}$. (Hint: Use the likelihood you obtained in Problem 3a. Then express π_1 and π_2 as in (1).) (2p)

b. How is the Fisher information matrix

$$\boldsymbol{J}(\alpha, \beta) = \begin{pmatrix} J_{11}(\alpha,\beta) & J_{12}(\alpha,\beta) \\ J_{21}(\alpha,\beta) & J_{22}(\alpha,\beta) \end{pmatrix}$$

derived from the second order derivatives of the log likelihood? (2p)

c. Compute the Fisher information matrix. (3p)

d. Use part 5c to show that

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) \\ \widehat{\mathrm{Cov}}(\hat{\beta}, \hat{\alpha}) & \widehat{\mathrm{Var}}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} \frac{1}{n_{11}} + \frac{1}{n_{12}} & -\frac{1}{n_{11}} - \frac{1}{n_{12}} \\ -\frac{1}{n_{11}} - \frac{1}{n_{12}} & \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \end{pmatrix}.$$

(Hint: In order to find the observed Fisher information matrix, you may use the ML-estimates of π_1 and π_2 of Problem 3c. You may also use without proof the formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

for the inverse of a 2×2 matrix.) (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