STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5019
EXAMINATION
February 20 2025

Categorical Data Analysis – Examination

February 20, 2025, 14:00-19:00

Examination by: Ola Hössjer, ph. 070 672 12 18, ola@math.su.se
Allowed to use: Miniräknare/pocket calculator and tables included in the appendix of this exam.
Återlämning/Return of exam: Announced through course homepage and web based course forum.

Each correct solution to an exercise yields 10 points.
Limits for grade: A, B, C, D, and E are 36, 32, 28, 24, and 20 points of 48 possible points (including 0-8 bonus 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

Patients with a previous heart attack were offered a new medicine in order to reduce the risk of new attacks. Before launching the medicine, the pharmaceutical company conducted a large pilot study in order to determine the impact that a weekly dose of x gram had on the risk

$$\pi(x) = P(Y = 1|x) = 1 - P(Y = 0|x)$$

of getting a new heart attack ($Y = 1$) within five years. They hypothesized that $\pi(x)$ followed a linear logistic regression model with parameters α and β within the range $x \in [0, 3]$ (no patient was given a higher weakly dose than 3 g).

a. Write down a formula for $\pi(x)$. (2p)

b. The maximum likelihood estimates of the parameters from the pilot study were $\hat{\alpha} = -1.5$ and $\hat{\beta} = -1.2$, with an estimated covariance matrix

$$\begin{pmatrix} \widehat{\text{Var}}(\hat{\alpha}) & \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}) \\ \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}) & \widehat{\text{Var}}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} 0.05 & -0.01 \\ -0.01 & 0.02 \end{pmatrix}.$$

Determine an approximate 95% confidence interval for the probability $\pi(1)$ of having a new heart attack within five years for patients whose weakly dose was 1 g. (Hint: Start finding a confidence interval for $\text{logit}[\pi(1)]$.) (4p)

c. Ben and Josh are two of the patients who had a previous heart attack. They both decided to take the new medicine, with weekly dosages of 1 g and 2.5 g respectively. (Not everyone took a high dose, since the medicine had known side effects.) Determine an approximate 95% confidence interval for the odds ratio between Ben and Josh of having a new heart attack within five years. (4p)

Problem 2

A certain toxic gas is known to increase the mortality rate. An epidemiologist defined a loglinear model, according to which the number of deaths $Y_i \sim \text{Po}(\mu_i)$ per year among individuals with i previous exposures to gas, are independent and Poisson distributed variables for $i = 0, 1, 2, 3$. The expected values

$$\mu_i = n_i \exp(\lambda_0 + \lambda_1 i), \quad i = 0, 1, 2, 3,$$

are proportional to the total number of individuals n_i with i previous exposures, whereas λ_0 and λ_1 are unknown parameters.

a. Define the log likelihood function $L(\lambda_0, \lambda_1)$ for data (y_0, y_1, y_2, y_3). (2p)

b. Find the likelihood equations. (Hint: Introduce the two score function components $u_j(\lambda_0, \lambda_1) = \partial L(\lambda_0, \lambda_1)/\partial \lambda_j$ for $j = 0, 1$.) (3p)

c. Find the Fisher information matrix $\boldsymbol{J}$ of the model in terms of the expected values μ_i. (Hint: Differentiate the score function components of b) with respect to λ_0 and λ_1 in order to find the 2×2 Hessian matrix of $L(\lambda_0, \lambda_1)$.) (3p)

d. Show that

$$\log(\boldsymbol{\mu}) = \begin{pmatrix} \log(\mu_0) \\ \log(\mu_1) \\ \log(\mu_2) \\ \log(\mu_3) \end{pmatrix} = \boldsymbol{X} \begin{pmatrix} \lambda_0 \\ \lambda_1 \end{pmatrix} + \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix},$$

for some appropriate model matrix $\boldsymbol{X}$ of dimension 4×2, and constants $c_0, \dots, c_3$. Then express $\boldsymbol{J}$ in terms of the model matrix and the expected values μ_i. (2p)

Problem 3

Multiple sclerosis (MS) is a neurological disease with genetic and environmental risk factors. It is known that smoking and allele 15 of the HLA-DRB1 gene both increase the risk of developing MS. A medical lab registered presence/absence of allele 15, disease status and smoking habits for a number of individuals, as shown in the table below. Let N_{ijk} refer to the number of observations with $X = i$, $Y = j$ and $Z = k$. Assume that these cell counts $N_{ijk} \sim \text{Po}(\mu_{ijk})$ are independent and Poisson distributed random variables.

Full table:

Allele 15?	Non-smokers $Z = 0$		Smokers $Z = 1$	
	No MS $Y = 0$	MS $Y = 1$	No MS $Y = 0$	MS $Y = 1$
No ($X = 0$)	154	116	20	30
Yes ($X = 1$)	61	69	15	35
Total (n_{+jk})	215	185	35	65

Marginal table for X, Y:

	$Y = 0$	$Y = 1$
$X = 0$	174	146
$X = 1$	76	104
Total (n_{+j+})	250	250

a. Determine the parameters of the loglinear model $M = (XY, ZY)$ and express μ_{ijk} in terms of these parameters. Specify which of the parameters you put to zero. (2p)

b. Prove that $\mu_{ijk} = \mu_{ij+}\mu_{+jk}/\mu_{+j+}$ for model M in a). (2p)

c. Use b) and data from the full and marginal contingency tables above in order to find the fitted expected cell counts $\hat{\mu}_{ijk}$ of model M. (3p)

d. Determine if M provides a good fit by computing the deviance

$$G^2(M) = 2\sum_{ijk} n_{ijk} \log \frac{n_{ijk}}{\hat{\mu}_{ijk}},$$

and conclude whether M should be rejected at level 5% or not. (3p)

Problem 4

We continue to analyze the dataset of Problem 3, regarding Y as the outcome variable, and X, Z as predictors.

a. Prove that the loglinear model $M = (XY, YZ)$ gives rise to a logistic regression model for $P(Y = 1|X = i, Z = k)$, and find the parameters of this model. Specify which of them you put to zero if $i = 0$ and $k = 0$ are used as baseline levels. (3p)

b. Express the three odds ratios

$$\theta_{ik} = \frac{P(Y = 1|X = i, Z = k)/(1 - P(Y = 1|X = i, Z = k))}{P(Y = 1|X = 0, Z = 0)/(1 - P(Y = 1|X = 0, Z = 0))}$$

for $(i, k) \in \{(0, 1), (1, 0), (1, 1)\}$ in terms of the parameters you found in a). (3p)

c. Let $\boldsymbol{\theta}$ be the parameters you found in a), and $\hat{\boldsymbol{\theta}}$ an estimate obtained by maximizing the prospective likelihood

$$l(\boldsymbol{\theta}) = \prod_{i,k=0}^{1} P(N_{i0k} = n_{i0k}, N_{i1k} = n_{i1k} | X = i, Z = k). \tag{1}$$

This likelihood is based on the assumption that data is drawn from the population distribution of $Y|X, Z$. But the data set of Problem 3 is actually sampled from the population distribution of $X, Z|Y$; a case-control study with 250 cases drawn randomly from all individuals affected by MS; and 250 controls drawn randomly from all healthy individuals. The likelihood in (1) is therefore misspecified. However, in spite of this, motivate why two of the parameters in $\boldsymbol{\theta}$ could still be estimated consistently by the corresponding components of $\hat{\boldsymbol{\theta}}$ if more data were collected, so that the number of cases and controls would increase. (Hint: You don't need to look at the likelihood equations. Use Bayes' Theorem instead to express the odds ratios from b) in terms of the $X, Z|Y$-distribution. Then make a qualitative argument what impact this has on estimating the components of $\boldsymbol{\theta}$.) (4p)

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