STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5019
EXAMINATION
February 13 2024

Categorical Data Analysis – Examination

February 13, 2024, 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 at the appendix of this exam.
Grading: 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 I refer to the annual income of a randomly chosen person of a large population, and $\mu = E(I)$ the expected annual income. A statistician investigated the effect that the base 2 logarithm $X = \log_2(I/\mu)$ of income (relative to population average) has on the probability that a person gets bankrupt ($Y = 1$) or not ($Y = 0$) within a five year period. He designed a cohort study with 10000 randomly chosen individuals. At first, each participant reported his or her income salary. Then five years later it was checked whether they had any bankruptcy. The estimated intercept and slope parameters of a logistic regression model where $\hat{\alpha} = -4.7$ and $\hat{\beta} = -0.85$. They are jointly approximately normally distributed, with an estimated covariance matrix

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) \\ \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) & \widehat{\mathrm{Var}}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} 0.015 & -0.003 \\ -0.003 & 0.005 \end{pmatrix}.$$

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

b. What is the predicted probability $\pi(-3)$ of bankruptcy for a person whose annual income is eight times lower than average? (2p)

c. Compute a 95% confidence interval for $\pi(-3)$, by first constructing a confidence interval for $\text{logit}[\pi(-3)] = \alpha - 3\beta$. (Hint: The 97.5% percentile of a standard normal distribution is 1.96.) (4p)

d. Compute a 95% confidence interval for the odds ratio of bankruptcy between Adam and Ben, if Adam's income is twice as large as Ben's, by first computing the corresponding confidence interval for the log odds ratio. (Hint: If the value of Adam's predictor variable is x_1, then the value of Ben's predictor variable is $x_2 = x_1 - 1$.) (3p)

Problem 2

After completing the study of Problem 1, the statistician wanted to confirm the results by collecting new data. He confined himself to a small subpopulation that consisted of two subgroups, those with annual income close to μ and 2μ respectively. Within this subpopulation, he randomly collected 500 cases, individuals with bankruptcy the last five years, and 500 controls with no bankruptcy, and for all of them he registered whether they belonged to the high or low income group 2μ or μ. He summarized data in terms of a 2×2 contingency table with income classes low ($X = 0$) and high ($X = 1$) as rows, whereas no bankruptcy ($Y = 0$) and bankruptcy ($Y = 1$) served as columns. He wanted to compare the odds of a high income individual between cases and controls, in terms of an odds ratio $\text{OR}^* = \exp(\beta^*)$. This analysis gave $\hat{\beta}^* = -0.76$ and $\widehat{\text{Var}}(\hat{\beta}^*) = 0.007$.

a. What kind of sampling scheme was actually used? Give your answer in terms of columns/rows and an appropriate distribution for each column/row. (2p)

b. Define the odds ratio OR^* in terms of probabilities $P(X = i|Y = j)$. (2p)

c. The statistician wanted to test

$$\begin{aligned} H_0 &: \ \beta = \beta^*, \\ H_a &: \ \beta \neq \beta^*, \end{aligned}$$

i.e. if the effect parameters of the two studies in Problems 1 and 2 were the same (to check if it was justified to pool the studies). He used a two-sided Wald test with 5% significance level to test if $\hat{\beta} - \hat{\beta}^*$ is significantly different from 0. Perform this test, assuming that the two parameter estimates $\hat{\beta}$ and $\hat{\beta}^*$ are independent, so that $\text{Var}(\hat{\beta} - \hat{\beta}^*) = \text{Var}(\hat{\beta}) + \text{Var}(\hat{\beta}^*)$. (Hint: Make use of the estimated variances of $\hat{\beta}$ and $\hat{\beta}^*$ from Problems 1 and 2, and that the 97.5% percentile of a standard normal distribution is 1.96.) (3p)

d. Show that the null hypothesis $\beta^* = \beta$ in 2c is actually correct if the logistic regression model of Problem 1 holds in Problem 2 as well. (Hint: Use $\text{OR}^* = \exp(\beta^*)$, Problem 2b and Bayes' Theorem.) (3p)

Problem 3

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 X and Y are jointly independent of Z. 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 Problem 3a to prove that

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

where a plus sign denotes summation over the corresponding index. (Hint: Use 3a to write $\mu_{ijk} = A_{ij}B_k$ for some appropriate factors A_{ij} and B_k.) (2p)

c. Use 3b 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 compute the observed values n_{ij+} for the marginal table of X and Y. The total sizes of the two partial tables are $n_{++1} = 174$ and $n_{++2} = 86$.) (2p)

Observed values n_{ij1}:

	$j=1$	$j=2$
$i=1$	65	42
$i=2$	29	38

Observed values n_{ij2}:

	$j=1$	$j=2$
$i=1$	20	32
$i=2$	19	15

d. Test the null hypothesis $M = (XY, Z)$ against the saturated model (XYZ) using a chisquare test with test statistic $X^2(M)$ and significance level 0.05. (3p)

Problem 4

Different risk factors of type 2 diabetes were sought for in an epidemiological study. The investigators used an ANOVA type multiple logistic regression model, with response variable $Y = 1$ ($Y = 0$) for patients with (without) diabetes. The three predictor variables body mass index (X), lack of physical exercise (Z) and insulin concentration (W) were all categorized into three levels. A likelihood analysis was performed for different submodels M, with predictors (main effects and interactions of different orders) and deviance $G^2(M)$ reported for each model:

M	$G^2(M)$	$p(M)$
$(X*Z*W)$	0	
$(X*Z+X*W+Z*W)$	7.70	
$(X*Z+X*W)$	15.27	
$(X*Z+Z*W)$	31.76	
$(X*W+Z*W)$	20.43	
$(X+Z*W)$	36.11	
$(Z+X*W)$	24.57	
$(W+X*Z)$	38.61	
$(X+Z+W)$	41.57	
None	117.78	

It is assumed that all models (including the "None" model) contain an intercept α. Any model is balanced, so if it contains a certain interaction, all lower order interactions or main effects "within" this interaction are included as well. For instance, if the second order interaction $X*Z$ belongs to a model (with parameters β_{ik}^{XZ} for different levels $X=i, Z=k$), the β_i^X and β_k^Z main effect parameters of X and Z are included as well.

a. Write down a formula for $P(Y=1|X=i,Z=k,W=h)$ under submodel $(X*W+Z*W)$, with intercept, main effect parameters and interaction parameters. (Hint: Choose one value of X, Y, and Z respectively as a baseline level of each variable. Only main effects and interaction parameters without baseline indeces have nonzero values.) (2p)

b. How many parameters p does the model in 4a have? Motivate your answer. (1p)

c. Fill in the third column of the table and compute p for all models. (Hint: You don't have to explain all calculations in detail, but report the most important steps, using the reasoning in 4b as a template.) (2p)

d. Define AIC(M). Use the table to select the best model according the AIC criterion. (Hint: Minimizing AIC(M) is equivalent to minimizing $G^2(M)$ plus a penalty term that is a certain function of the number of parameters $p(M)$ of M.) (2p)

e. Suppose Forward Inclusion (FI) is used instead to select among the submodels of the table, with each hypothesis test at significance level 5%. Describe which pairs of models that are tested (using only those that are listed in the table), and which model that is eventually selected. (Hint: A likelihood ratio test statistic between two nested models is the difference between two deviances.) (3p)

Problem 5

An insurance company has a system with four bonus classes $i \in \{1,2,3,4\}$, where customers in higher classes had fewer accidents in the past and therefore pay a lower premium today. An actuary wants to find out whether the bonus class affects the current accident rate μ. She models the total number of reported accidents Y_i during one year, for drivers in different bonus classes $x_i = i$, as independent Poisson random variables

$$Y_i|x_i \sim \text{Po}(t_i\mu_i),$$

where t_i is the accumulated time of risk (in units of thousand years) for the drivers of class i, and $\mu_i = \exp(\alpha + \beta x_i)$. Then she collects the following data:

i	1	2	3	4
t_i	8.2	11.3	16.1	14.3
Y_i	502	760	921	630

a. Write down the log likelihood function $L(\alpha, \beta)$. (2p)

b. Compute the maximum likelihood estimate $\hat{\alpha}_0$ of α under the null hypothesis

$$H_0 : \beta = 0,$$

when testing whether bonus class affects current accident rate or not. (Hint: Make use of the likelihood score function component $u_1(\alpha, \beta) = \partial L(\alpha, \beta)/\partial\alpha$ of α when $\beta = 0$, and equate it to zero. Note also that $\partial\mu_i/\partial\alpha = \mu_i$ and $\partial \log(\mu_i)/\partial\alpha = 1$.) (2p)

c. In order to compute the maximum likelihood estimator $(\hat{\alpha}, \hat{\beta})$ of (α, β) under the full model, the actuary uses Newton-Raphson's iterative scheme, with $(\alpha^{(0)}, \beta^{(0)}) = (\hat{\alpha}_0, 0)$ as initial guess. Compute the improved approximation $(\alpha^{(1)}, \beta^{(1)})$ of $(\hat{\alpha}, \hat{\beta})$, after one iteration. (Hint: You will need to invert the 2×2 Hessian matrix of the log likelihood, and then use the formula

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

(4p)

d. Do the same as in 5c for the Fisher scoring iterative scheme. That is, compute the first iterate $(\alpha^{(1)}, \beta^{(1)})$ based on the same staring value $(\alpha^{(0)}, \beta^{(0)}) = (\hat{\alpha}_0, 0)$. (Hint: You don't need any numerical calculations. Only motivate whether the answer in 5d is different or not from that in 5c.) (2p)

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