STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5019
EXAMINATION
August 21 2025

Categorical Data Analysis – Examination

August 21, 2025, 8.00-13.00

Examination by: Ola Hössjer, ph. 08 16 45 84, ola@math.su.se
Allowed to use: Miniräknare/pocket calculator and tables included in 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 36, 32, 28, 24, and 20 points of 48 possible points (including bonus of 0-8 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

Lung cancer Y and the amount of smoking x were registered for 5000 individuals. Here Y is a binary variable with levels 0 and 1 corresponding to no cancer and cancer respectively, whereas $x \geq 0$ is measured in pack years. This number quantifies the cumulative amount of smoking, and it is obtained by multiplying the number of packs (of 20 cigarettes) a person has smoked daily, with the number of years of smoking. A linear logistic regression model with intercept α and effect parameter β was used to model the risk $\pi(x) = P(Y = 1|X = x)$ of developing lung cancer, for a patient with smoking habits x.

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

b. The maximum likelihood estimates of the parameters from the data set where $\hat{\alpha} = -2.2$ and $\hat{\beta} = 0.8$, with an estimated covariance matrix

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) \\ \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) & \widehat{\mathrm{Var}}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} 0.02 & -0.005 \\ -0.005 & 0.01 \end{pmatrix}.$$

Determine an approximate 95% confidence interval for the probability of developing lung cancer for an individual whose cumulative amount of smoking is 1.5 pack years. (4p)

c. Determine an approximate 95% confidence interval for the odds ratio of having lung cancer between Adam and Ben, both of which have smoked for one year, with Adam having 10 more cigarettes each day than Ben. (4p)

Problem 2

Now suppose that the sample of 5000 individuals from Problem 1 was not collected randomly, but rather that lung cancer patients where oversampled by a factor $\rho > 1$. This was done in order to obtain enough individuals with lung cancer in the sample. That is, within each group of individuals with the same smoking habit x, a sampled person's probability of lung cancer was

$$\pi^*(x) = P^*(Y = 1|X = x) = \frac{\rho P(Y = 1|X = x)}{P(Y = 0|X = x) + \rho P(Y = 1|X = x)},$$

rather than $\pi(x) = P(Y = 1|X = x)$.

a. Show that $\pi^*(x)$ defines another logistic regression model. Describe in particular how the intercept α^* and effect parameter β^* of this model relate to α and β of Problem 1. (3p)

b. Assume it is known that $\rho = 10$, and regard $\hat{\alpha}$ and $\hat{\beta}$ of Problem 1b) as unbiased estimates of α^* and β^*. Compute parameter estimates $\tilde{\alpha} = \tilde{\alpha}(\hat{\alpha}, \rho)$ and $\tilde{\beta}$ of α and β that take into account that lung cancer patients were oversampled according to model π^*. Then estimate the probability of having lung cancer for an individual whose cumulative amount of smoking is 1.5 pack years. Compare this estimate with the confidence interval that you found in Problem 1b, and draw a conclusion from this comparison. (3p)

c. Now suppose the value of ρ in b) was only an estimate $\hat{\rho} = 10$. This estimate of ρ was obtained by comparing the fraction of individuals with lung cancer, among the 5000 individuals in Problem 1, with the fraction of lung cancer in the whole population. It particular, variance and covariance estimates $\widehat{\text{Var}}(\hat{\rho}) = 0.5$ and $\widehat{\text{Cov}}(\hat{\alpha}, \hat{\rho}) = 0.005$ where computed, with $\hat{\alpha}$ the (biased) estimate of α from Problem 1. The corrected estimate $\tilde{\alpha}$ of α in b) can still be used, but in order quantify the variability of this estimate we need to take into account that ρ has been estimated. Use a two-dimensional first order Taylor expansion of the function $\tilde{\alpha} = \tilde{\alpha}(\hat{\alpha}, \hat{\rho})$ in order to find an approximation of $\text{Var}(\tilde{\alpha})$. Then find an estimator $\widehat{\text{Var}}(\tilde{\alpha})$ of this variance. (Hint: The value of $\widehat{\text{Var}}(\hat{\alpha})$ from Problem 1b) can be used.) (4p)

Problem 3

A group of epidemiologists investigated genetic components of Alzheimer's disease. They collected data from three binary variables; a previously known risk factor X, with levels

0 and 1 corresponding to two variants of a particular gene A; disease status Y, where unaffected and affected individuals had levels 0 and 1 respectively; and a new hypothesized risk factor Z, with levels 0 and 1 corresponding to two different variants of second gene B. The data set was summarized as a threeway table for X, Y, and Z. The number of observations N_{ijk} with $X = i$, $Y = j$ and $Z = k$ was assumed to be Poisson distributed with expected value μ_{ijk} for $0 \leq i, j, k \leq 1$, 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. This corresponds to a model where gene B has no effect on Alzheimer's disease. 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 a), or a direct argument, to prove that

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

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

c. Use b) and data n_{ijk} from the partial and marginal tables below to find the ML estimates $\hat{\mu}_{ijk}$ of all μ_{ijk}. (2p)

Observed values n_{ij0}:

	$j=0$	$j=1$
$i=0$	60	15
$i=1$	45	28

Observed values n_{ij1}:

	$j=0$	$j=1$
$i=0$	55	24
$i=1$	40	35

Observed values n_{ij+}:

	$j=0$	$j=1$
$i=0$	115	39
$i=1$	85	63

d. In order to test if one of the two variants of gene B is a risk factor for Alzheimer's disease, choose between $M = (XY, Z)$ and the saturated model (XYZ), using a chisquare test with test statistic $X^2(M)$ and significance level 0.05. (3p)

Problem 4

The actuaries of a non-life insurance company wanted to find out which factors best explained whether a customer experienced at least one accident ($Y = 1$) or not ($Y = 0$) during one year. They wanted to use an ANOVA type multiple logistic regression model with Y as outcome variable, and three categorical variables; annual mileage (A), type of car (Z), and type of region (W); as risk factors. Since the data set was large, they could afford having many parameters, and therefore they chose to have 4 different levels of each risk factor. On the other hand, since mileage A had highest priority, they did not include interaction between type of car and region in any submodel M. A likelihood analysis was performed for different M, with the following predictors (main effects and second order interactions) and log likelihoods $L(M)$:

M	$L(M)$	$p(M)$
$(A*Z + A*W)$	-5000.0	
$(Z + A*W)$	-5005.5	
$(W + A*Z)$	-5008.0	
$(A + Z + W)$	-5025.0	
None	-5200.0	

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 main effects within this interaction are included as well. For instance, if the second order interaction $A * W$ belongs to a model (with parameters β_{ih}^{AW} for different levels $A = i, W = h$), the main effect parameters β_i^A and β_h^W, of A and W respectively, are included as well.

a. Write down a formula for $P(Y = 1|A = i, Z = k, W = h)$ under submodel $(A * Z + A * W)$, with intercept, main effect parameters, and interaction parameters. (2p)

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

c. Compute p for all models, i.e. the third column of the table. (Hint: You don't have to explain all calculations in detail. Report the most important steps, using the reasoning in b) as a template.) (2p)

d. Define AIC(M). Use the table to select the best model according the AIC criterion. (2p)

e. Suppose Backward Elimination (BE) is used instead to select among the submodels of the table, with each hypothesis tested 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 eventually is selected. (3p)

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