STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5006
EXAMINATION
February 14 2020

Categorical Data Analysis – Examination

February 14, 2020, 9.00-14.00

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: 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 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 1

It is well known that men with prostate cancer often have an elevated concentration of a certain protein (Prostate-Specific Antigen, PSA) in their blood. A group of epidemiologists wanted to find out how effective, as a predictor of prostate cancer, the PSA-level x is. To this end they measured x (in units of nanograms per milliliter, ng/mL) and disease status ($Y = 1$ if affected, $Y = 0$ otherwise) among 10,000 men. 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 having prostate cancer, for a man with PSA-level x.

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

b. The maximum likelihood estimates of the parameters from the data set where $\hat{\alpha} = -6$ and $\hat{\beta} = 0.5$, with an estimated covariance matrix

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) \\ \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) & \widehat{\mathrm{Var}}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} 0.1 & -0.01 \\ -0.01 & 0.005 \end{pmatrix}.$$

Determine an approximate 95% confidence interval for the probability of having prostate cancer for a man with a PSA-level of 8. (4p)

c. Determine an approximate 95% confidence interval for the odds ratio of having prostate cancer between James and Scott, if James' PSA-level is 3 units higher than Scott's. (4p)

Problem 2

The table below shows the outcome of two type of heart surgeries, A and B. For each person it was registered whether there was any complication after the surgery or not. Denote by $\{n_{ij};\, 1 \leq i, j \leq 2\}$ the cell counts of the table, where i is the row number and j the column number. Regard this as Poisson sampling, so that n_{ij} are observations of independent Poisson variables $N_{ij} \sim \text{Po}(\mu_{ij})$.

	Outcome of surgery	
Type of surgery	1: Complication	2: No complication
1: A	30	170
2: B	20	80

a. Let π_1 (π_2) be the probability that a person who went through a surgery of type A (B) had a complication. Express these two probabilities in terms of the expected cell counts μ_{ij}. (2p)

b. What is the joint distribution of N_{11} and N_{21} when one conditions on the two row sums $N_{1+} = n_1 = n_{1+}$ and $N_{2+} = n_2 = n_{2+}$? (Hint: Start by defining the marginal distributions of N_{11} and N_{21}, given their respective row sums n_1 and n_2.) (2p)

c. Introduce an appropriate estimator $\hat{r} = \hat{\pi}_1/\hat{\pi}_2$ of the relative risk $r = \pi_1/\pi_2$. Assume the sampling scheme of b) and use the multivariate delta method (i.e. a Taylor expansion of $\log(\hat{r})$ with respect to $\hat{\pi}_1$ and $\hat{\pi}_2$) to prove that

$$\text{Var}\left[\log(\hat{r})\right] \approx \frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2}.$$

(3p)

d. Use c) to find an approximate 95% two-sided confidence interval for r. Conclude from this whether or not type of surgery had a significant effect at level 5% on the probability of a complication. (Hint: Start by computing a confidence interval for $\log(r)$.) (3p)

Problem 3

A survey was performed in Ohio among 2276 high school students. They were asked whether they ever used alcohol (A), cigarettes (C) or marijuana (M):

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Let N_{acm} refer to the number of individuals with $A = a$, $C = c$, and $M = m$, assuming (say) that the three binary variables are encoded as 0 for non-users and 1 for users. It is further assumed that all $N_{acm} \sim \text{Po}(\mu_{acm})$ are independent and Poisson distributed random variables. A number of balanced loglinear models are used to describe how the expected cell counts μ_{acm} depend on A, C, M, and fitted to data in terms of their deviances G^2(Model). This is shown in the following table:

Model	G^2(Model)	p(Model)
(A, C, M)	1286.0	
(A, CM)	534.2	
(C, AM)	939.6	
(M, AC)	843.8	
(AC, AM)	497.4	
(AC, CM)	92.0	
(AM, CM)	187.8	
(AC, AM, CM)	0.4	
(ACM)	0.0	

a. Express μ_{acm} in terms of loglinear parameters for (AC, AM, CM). Discuss in particular which loglinear parameters you put to zero in order to avoid overparametrization. Then compute the number of parameters $p(AC, AM, CM)$. (3p)

b. Compute the number of parameters p(Model) for all models in the table above. (Hint: You don't have to define the parameters for all these model. It suffices to describe how you obtain p(Model).) (2p)

c. Define Akaike's information criterion AIC(Model) for a model in terms of the log likelihood, and select the best model according to this criterion, among those listed in the table. (Hint: You don't need to know AIC(Model) for any model in order to select the best one. The information from the table above is sufficient.) (2p)

d. Select the best model using backward elimination (BE) (with significance level 5% for all tests), among those listed in the table. That is, start with the largest model among those that appear in the table. (3p)

Problem 4

A medical doctor regards marijuana as a more serious health threat than alcohol and cigarettes. For the data set of Problem 3, he therefore treated M as the outcome variable and A, C as predictor variables of marijuana use.

a. For the loglinear model (AC, AM, CM) of Problem 3a), show that $P(M = 1|A = a, C = c)$ defines an ANOVA type multiple logistic regression model. Express the parameters of this model as functions of the loglinear parameters in Problem 3a), and determine how many of the logistic regression parameters that are nonzero. (2p)

b. Define the conditional odds ratio $\theta_{AM(c)}$ of using marijuana for those that use alcohol, compared to those that don't, for individuals with smoking habit c. Then express $\theta_{AM(c)}$ for the logistic regression model in 4a), first in terms of the logistic regression parameters, and then in terms of the loglinear parameters in 3a). (3p)

c. Define what homogeneous association between M and A means. Which models in the table of Problem 3 have homogeneous association between M and A? (Hint: You don't have to compute $\theta_{AM(c)}$ for all models. A general argument is sufficient.) (2p)

d. Consider the loglinear model (AM, AC). Prove that this model not only has homogeneous association, but also that the conditional odds ratio $\theta_{AM(c)}$ equals the marginal odds ratio θ_{AM} between alcohol and marijuana use. In particular, compute the maximum likelihood estimator $\hat{\theta}_{AM}$ of θ_{AM} from the data set of Problem 3. Comparing this estimate with $\hat{\theta}_{AM(c)} = 19.8$ for model (AC, AM, CM) of Problem 3a), which one do you think is most reliable? (Hint: Use the fact that C and M are conditionally independent given A for model (AM, AC).) (3p)

Problem 5

A group of epidemiologists wanted to find out whether previous occurrences of a certain type of heart attack increased the risk of getting new ones. To this end they defined a loglinear model, according to which the total number of heart attacks $Y_i \sim \text{Po}(\mu_i)$ among patients with i previous attacks, are independent and Poisson distributed variables for $i = 0, 1, 2, 3$. The expected values

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

are proportional to the the total time t_i at risk for individuals with i previous heart attacks, whereas λ_0 and λ_1 are unknown parameters. The objective of the study was to test the null hypothesis $H_0 : \lambda_1 = 0$ against the alternative hypothesis $H_a : \lambda_1 > 0$.

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. (3p)

c. Find the Fisher information matrix $\boldsymbol{J}$ of the model in terms of the expected values μ_i. (3p)

d. Suppose $\hat{\lambda}_0$ and $\hat{\lambda}_1$ are the maximum likelihood estimators of λ_0 and λ_1. Define the Wald test (with approximate significance level α) for testing H_0 against H_a. (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