STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics

MT 5022
EXAMINATION
February 11 2026

# Classification and Analysis of Categorical Data – Examination

February 11, 2026, 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 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 44 possible points (including bonus of 0-4 points from computer assignments).

Reasoning and notation should be clear. You might answer in Swedish or English.

Read through the whole exam at first. Exercises need not to be ordered from simpler to harder.

---

# Problem 1

A toxic gas, which by mistake was emitted from a factory, caused a number of deaths among people who lived in the surrounding area. The gas was inhaled in the lungs and absorbed into the blood, and from there it spread to a number of tissues. A group of epidemiologists wanted to find out how the blood concentration (partial pressure in units of mm Hg) of the gas affected the risk of death. They formulated a quadratic logistic regression model

$$P(Y=1|X=x)=\pi(x)=\frac{\exp(\beta_0+\beta_1 x+\beta_2 x^2)}{1+\exp(\beta_0+\beta_1 x+\beta_2 x^2)} \tag{1}$$

for the probability that an individual with blood concentration $x$, died within one month after the gas was emitted. Blood samples were taken from 1000 randomly chosen people

the day after the release of the toxic gas. After fitting model (1) to data, the epidemiologists found that the maximum likelihood estimates of the parameters were $\hat{\beta}_0 = -6.0$, $\hat{\beta}_1 = 1.0$, and $\hat{\beta}_2 = 0.5$, and the covariance matrix of these parameter estimates were estimated as

$$\begin{pmatrix} \widehat{\text{Var}}(\hat{\beta}_0) & \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) & \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_2) \\ \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_0) & \widehat{\text{Var}}(\hat{\beta}_1) & \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) \\ \widehat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_0) & \widehat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_1) & \widehat{\text{Var}}(\hat{\beta}_2) \end{pmatrix} = \begin{pmatrix} 0.01 & -0.01 & -0.01 \\ -0.01 & 0.02 & -0.01 \\ -0.01 & -0.01 & 0.02 \end{pmatrix}.$$

a. Compute a 95% confidence interval for the death probability $\pi(2)$ of a person with gas concentration 2 mm Hg in the blood, by first finding the corresponding 95% confidence interval of $\text{logit}[\pi(2)]$. (5p)

b. Find an expression for the logarithm of the odds ratio of dying, between two individuals with gas concentrations 2 mm Hg and 1 mm Hg respectively, in terms of the logistic regression parameters. (2p)

c. Use b) to compute a 95% confidence interval for the odds ratio of dying, between two individuals with gas concentrations 2 mm Hg and 1 mm Hg respectively, by first finding the 95% confidence interval for the corresponding log odds ratio. (3p)

# Problem 2

A group of geneticists, who worked at a governmental medical agency, had recently found two genes I and II of importance for metabolism. More specifically, they discovered two variants $a$ and $A$ of I, and two variants $b$ and $B$ of II, where both the $A$ variant and the $B$ variant up-regulated metabolism. Since I and II are both located on non-sex chromosomes, each individual has $X \in \{0, 1, 2\}$ copies of $A$ and $Z \in \{0, 1, 2\}$ copies of $B$. The researchers wanted to find out whether presence of $A$ in an individual was correlated with presence of $B$ or not, because this would convey information as to whether the two proteins that I and II coded for, were involved in the same metabolic pathway or not. To this end, they selected a random sample of 500 individuals from the population, and registered $X$ and $Z$ for each one of them. They modeled this as multinomial sampling for all cell counts $\{N_{ik};\, 0 \leq i, k \leq 2\}$, with $\pi_{ik} = P(X = i, Z = k)$ the probability of having $X$ and $Z$ at levels $i$ and $k$. The result of the study is summarized in the following table:

| | $Z$ | | | |
|---|---|---|---|---|
| $X$ | 0 | 1 | 2 | Sum |
| 0 | 30 | 60 | 28 | 118 |
| 1 | 47 | 120 | 89 | 256 |
| 2 | 16 | 60 | 50 | 126 |
| Sum | 93 | 240 | 167 | 500 |

a. Formulate the likelihood in terms of 8 parameters for the saturated multinomial model. (Hint: The cell probabilities $\pi_{ik}$ sum to 1.) (2p)

b. Use a $X^2$-test statistic in order to test the null hypothesis $H_0 : \pi_{ik} = \pi_{i+}\pi_{+k}$ against the saturated model, which corresponds to an alternative hypothesis $H_a$ that $\pi_{ik} \neq \pi_{i+}\pi_{+k}$ for at least one cell $(i, k)$. Is $H_0$ rejected at significance level 5%? (Hint: The estimate of the expected count $\mu_{ik} = E(N_{ik})$ of cell $(i, k)$ under $H_0$ is $n\hat{\pi}_{i+}\hat{\pi}_{+k}$, where $n = N_{++} = 500$ is the total cell count, whereas $\hat{\pi}_{i+}$ and $\hat{\pi}_{+k}$ are estimates of $\pi_{i+}$ and $\pi_{+k}$.) (4p)

c. The researchers wanted to modify the null hypothesis of the test, so that *not only* $A$ and $B$ occurred independently, but *also* that each individual's two copies of $A$ occurred independently, i.e.

$$P(X = i) = \pi_{i+} = \begin{cases} (1-p)^2, & i = 0, \\ 2p(1-p), & i = 1, \\ p^2, & i = 2, \end{cases} \tag{2}$$

for some $p = P(A)$, *and* similarly that both copies of $B$ occurred independently, so that

$$P(Z = k) = \pi_{+k} = \begin{cases} (1-q)^2, & k = 0, \\ 2q(1-q), & k = 1, \\ q^2, & k = 2, \end{cases} \tag{3}$$

for some $q = P(B)$. Perform a new $X^2$-test of the new null hypothesis $H_0'$ (where the extra requirements (2) and (3) are added to $H_0$ in b)) against the saturated model. Is $H_0'$ rejected at level 5%? (Hint: There are two free parameters $p$ and $q$ under $H_0'$. Start by estimating $p$ and $q$ from the marginal cell counts of $X$ and $Z$. Then estimate $\mu_{ik}$ under $H_0'$ by $n\hat{\pi}'_{i+}\hat{\pi}'_{+k}$, where $\hat{\pi}'_{i+}$ and $\hat{\pi}'_{+k}$ are estimates of $\pi_{i+}$ and $\pi_{+k}$ that involve the estimates of $p$ and $q$ respectively.) (4p)

# Problem 3

Seat belt use $S$, injury $I$, location $L$ and gender $G$ were reported for 68694 passengers involved in automobiles accidents in the state of Maine. The dataset is summarized in the following $2 \times 2 \times 2 \times 2$ contingency table:

| | | Seat | Injury? | |
|---|---|---|---|---|
| Gender | Location | Belt? | No | Yes |
| Female | Urban | No | 7287 | 996 |
| | | Yes | 11587 | 759 |
| | Rural | No | 3246 | 973 |
| | | Yes | 6134 | 757 |
| Male | Urban | No | 10381 | 812 |
| | | Yes | 10969 | 380 |
| | Rural | No | 6123 | 1084 |
| | | Yes | 6693 | 513 |

Let $N_{gils}$ refer to the number of individuals with $G = g$, $I = i$, $L = l$, and $S = s$, assuming (say) that the four binary variables are encoded as 0=female, 1=male, 0=no injury, 1=injury, 0=urban, 1=rural, 0=no seat belt use, and 1=seat belt use. It is assumed

that all $N_{gils} \sim \text{Po}(\mu_{gils})$ are independent and Poisson distributed random variables. A number of balanced loglinear models $M$ are used to describe how the expected cell counts $\mu_{gils}$ depend on $G, I, L, S$, and fitted to data in terms of their deviances $G^2(M)$. This is shown in the following table:

| Model $M$ | $G^2(M)$ | $p(M)$ |
|---|---|---|
| $(G, I, L, S)$ | 2792.8 | |
| $(GI, GL, GS, IL, IS, LS)$ | 23.4 | |
| $(GIL, GS, IS, LS)$ | 18.6 | |
| $(GIS, GL, IL, LS)$ | 22.8 | |
| $(GLS, GI, IL, IS)$ | 7.5 | |
| $(ILS, GI, GL, GS)$ | 20.6 | |
| $(GIL, GIS, GLS, ILS)$ | 1.33 | |

a. Express $\mu_{gils}$ in terms of loglinear parameters for $M = (GI, GL, GS, IL, IS, LS)$. Discuss in particular which loglinear parameters you put to zero in order to avoid overparametrization, and compute $p(M)$, the number of parameters of $M$. (3p)

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

c. Define Akaike's information criterion AIC($M$) for model $M$ 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($M$) for any model $M$ in order to select the one with minimal AIC($M$). The information from the three columns of 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

For the accidents data set of Problem 3, let $I$ be the outcome variable and $G, L, S$ the predictor variables.

a. For model $M = (GI, GL, GS, IL, IS, LS)$ of Problem 3a), show that $P(I = 1|G = g, L = l, S = s)$ 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 $\theta_{IS(gl)}$ of having an injury for those that use seat belt, compared to those that don't, for individuals of gender $g$ that live in region $l$. Then express $\theta_{IS(gl)}$ for model $M$, first in terms of the logistic regression parameters in 4a), then in terms of the loglinear parameters in 3a). (Hint: Start looking at $\log(\theta_{IS(gl)})$.) (3p)

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

d. Now consider the loglinear model $M_0 = (IS, IGL)$. Prove that $M_0$ not only has homogeneous association, but also that the conditional odds ratio $\theta_{IS(gl)}$ for $M_0$ equals the marginal odds ratio $\theta_{IS}$ between injury and seat belt use. In particular, compute the maximum likelihood estimator $\hat{\theta}_{IS}$ of $\theta_{IS}$ from the data set of Problem 3 and compare this estimate with $\hat{\theta}_{IS(gl)} = 0.44$ for model $M = (GI, GL, GS, IL, IS, LS)$. (Hint: Use Bayes' Theorem and the fact that $S$ and $(G, L)$ are conditionally independent given $I$.) (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 |