UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Analysis of Categorical Data 1MS370
2020–08–28
Solutions

Time: 8.00-13.00. Limits for the credits 3, 4, 5 are 18, 25 and 32 points, respectively. The solutions should be well motivated.

1. A discrete random variable Y has probability mass function

$$p(y;\pi) = \pi^{3y}(1-\pi^3)^{1-y},$$

for $y = 0, 1$. Assume that $0 < \pi < 1$.

(a) Does this distribution belong to the exponential family, and in that case, why? (2p)

Solution: We may write

$$p(y;\pi) = (1-\pi^3)\exp\left\{y\log\left(\frac{\pi^3}{1-\pi^3}\right)\right\},$$

and so, this is an exponential family mass function $a(\pi)b(y)\exp\{yQ(\pi)\}$ with

$$a(\pi) = 1-\pi^3, \quad b(y) = 1, \quad Q(\pi) = \log\left(\frac{\pi^3}{1-\pi^3}\right),$$

where $Q(\pi)$ is the natural parameter.

(b) Suggest an appropriate link function $g(\pi)$. (2p)

Solution: The appropriate link function is the natural parameter, i.e.

$$g(\pi) = \log\left(\frac{\pi^3}{1-\pi^3}\right).$$

(c) Let x be an explanatory variable that can take any real value. Discuss if the GLM $g(\pi) = \alpha + \beta x$ is a suitable model. (2p)

Solution: As π goes from 0 to 1, $g(\pi)$ increases monotonically from $-\infty$ to ∞. Also, $\alpha + \beta x$ may in principle take any real value. Hence, we have a suitable model.

2. In the 2000 general Society Syrvey in the US, a random sample of voters were asked about their political party identification. The results are given in the table below.

	Democrat	Independent	Republican
Females	762	327	468
Males	484	239	477

(a) Test if political party identification was independent of gender. (3p)

Solution: Test H_0: political party identification was independent of gender vs H_1: $\neg H_0$ (dependence).

The row sums are 1557 and 1200, and the column sums are 1246, 566 and 945. In total, there are 2757 observations. This gives the expected cell counts

$$e_{11} = \frac{1557 * 1246}{2757} = 703.67,\ e_{12} = \frac{1557 * 566}{2757} = 319.65,\ e_{13} = \frac{1557 * 945}{2757} = 533.68,$$
$$e_{21} = \frac{1200 * 1246}{2757} = 542.33,\ e_{22} = \frac{1200 * 566}{2757} = 246.35,\ e_{23} = \frac{1200 * 945}{2757} = 411.32.$$

These are all greater than 5, hence χ^2 approximation is permitted. We get

$$\begin{aligned} X^2 &= \frac{(762 - 703.67)^2}{703.67} + \frac{(327 - 319.65)^2}{319.65} + \frac{(468 - 533.68)^2}{533.68} \\ &\quad + \frac{(484 - 542.33)^2}{542.33} + \frac{(239 - 246.35)^2}{246.35} + \frac{(477 - 411.32)^2}{411.32} \\ &\approx 30.1. \end{aligned}$$

The number of degrees of freedom is $(3-1)\cdot(2-1) = 2$. We have $\chi^2_{0.05}(2) = 5.99$, and since $30.1 > 5.99$, we may reject H_0 at the 5% level. (We may even reject H_0 at the 0.1% level, because $\chi^2_{0.001}(2) = 13.82$.)

Hence, there is a significant dependence between party identification and gender.

(b) Partition the G^2 statistic to see if the gender comparison as in (a) turns out different when considering only Democrat vs Independent and then Democrat or Independent vs Republican. (3p)

Solution: The G^2 statistic for independence for the whole table is

$$\begin{aligned} G^2 &= 2 \cdot \left\{ 762 \log\left(\frac{762}{703.67}\right) + 327 \log\left(\frac{327}{319.65}\right) + 468 \log\left(\frac{468}{533.68}\right) \right. \\ &\quad \left. +484 \log\left(\frac{484}{542.33}\right) + 239 \log\left(\frac{239}{246.35}\right) + 477 \log\left(\frac{477}{411.32}\right) \right\} \\ &\approx 30.01. \end{aligned}$$

Taking the sub table with only the Democrat and Independent columns, we have the row sums 1089 and 723, and as before, the column sums are 1246 and 566. The total number of observations is 1812, and the expected cell counts are

$$\begin{aligned} e_{11} &= \frac{1089 * 1246}{1812} = 748.84, \; e_{12} = \frac{1089 * 566}{1812} = 340.16, \\ e_{21} &= \frac{723 * 1246}{1812} = 497.16, \; e_{22} = \frac{723 * 566}{1812} = 225.84. \end{aligned}$$

These numbers are all greater than 5, so χ^2 approximation is permitted. The G^2 statistic for this sub table is

$$\begin{aligned} G_1^2 &= 2 \cdot \left\{ 762 \log\left(\frac{762}{748.84}\right) + 327 \log\left(\frac{327}{340.16}\right) \right. \\ &\quad \left. +484 \log\left(\frac{484}{497.16}\right) + 239 \log\left(\frac{239}{225.84}\right) \right\} \\ &\approx 1.85. \end{aligned}$$

Interestingly, we have only one degree of freedom and $\chi^2_{0.05}(1) = 3.84$, so at the 5% level, the null hypothesis of independence is not rejected vs the alternative of dependence for this sub table.

As for testing if Democrat or Independent vs Republican is independent of gender, we may use the corresponding G^2 statistic

$$G_2^2 = G^2 - G_1^2 = 30.01 - 1.85 = 28.16,$$

which compared to $\chi^2_{0.001}(1) = 10.83$ gives a significant result at the 0.1% level.

Apparently, the gender dependence only comes into play when moving from Independent to Republican in the table.

3. Let $P(Y = 1|x) = \pi(x) = F(\alpha + \beta x)$. Consider the following suggestions for the function F (below, $\pi \approx 3.14$, not a probability):

   (*i*) $F(z) = \begin{cases} 0 & \text{if } z < 0, \\ \sin\left(\frac{\pi}{2}z\right) & \text{if } 0 \leq z \leq 1, \\ 1 & \text{if } z > 1, \end{cases}$

   (*ii*) $F(z) = \frac{1}{\pi}\arctan(z) + \frac{1}{2}$,

   (*iii*) $F(z) = \frac{2}{\pi}\arctan(z)$.

   (a) Which (if any) of the suggestions gives a suitable model, and which do not? Why or why not? (2p)

   Solution: A suitable $F(z)$ should behave like a cumulative distribution function (cdf), i.e. it should be monotone in the argument z and map all real z on the unit interval $(0, 1)$ (because $\pi(x)$ is a probability). All suggested functions are monotonely non decreasing, so that is fine. Functions (*i*) and (*ii*) map to the unit interval, but not (*iii*) which can take negative values. So in this sense both (*i*) and (*ii*) are suitable.

   However, (*ii*) might be considered as the best alternative, since it maps the argument z on the unit interval in a smooth way, while (*i*) maps all negative arguments to zero and all arguments above one to one.

   (b) Take your favourite choice of function F from above. For which x (as a function of α and β) is it true that the function $\pi(x) = 1/2$? (2p)

   Solution: Let as take function (*ii*). At first, we solve

$$\frac{1}{2} = F(z) = \frac{1}{\pi}\arctan(z) + \frac{1}{2}$$

   for z, which gives $z = 0$. Then, $\alpha + \beta x = 0$ gives $x = -\alpha/\beta$.

   (c) What is the rate of increase of the function $\pi(x)$ at this point? (2p)

   Solution: Take $z = \alpha + \beta x$. The rate of increase is given by the derivative (use the chain rule)

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial z}\frac{\partial z}{\partial x} = \frac{1}{\pi}\frac{1}{1 + z^2}\beta,$$

   which at $z = 0$ is β/π.

4. The grades for those who passed the course "Applied Statistics" in March 2020 at the STS program in Uppsala, for year of birth -97 (up to 1997) and 98- (1998 and on) are given in the table below.

   All probabilities mentioned below should be interpreted as conditional probabilities given that the student has passed the exam.

	-97	98-
3	11	2
4	16	14
5	12	10

   (a) Based on the table, estimate the ratio of the probability to get a grade of 4 to the probability to get a grade of 3 for the two age groups. (1p)

   Solution: There were 39 passing students born -97. Hence, the probability for such a student to get a grade of 4 was 16/39, and the probability to get a grade of 3 was 11/39. It follows that the required ratio for students born -97 is

$$\frac{16/39}{11/39} = \frac{16}{11} \approx 1.4545,$$

   and similarly, for students born 98- it is

$$\frac{14}{2} = 7.$$

   (b) Consider the baseline-category model

$$\log\left\{\frac{\pi_j(x)}{\pi_1(x)}\right\} = \alpha_j + \beta_j x,$$

   where $j = 2, 3, ..., J$ and $\pi_j(x) = P(Y = j|x)$ with $x = 0$ for born -97 and 1 for born 98-.

   Such a model was estimated based on the data, where $j = 1, 2, 3$ correspond to the grades 3, 4, 5, respectively.

   The parameter estimates were $\hat{\alpha}_2 = 0.37469$, $\hat{\alpha}_3 = 0.08701$, $\hat{\beta}_2 = 1.57122$, $\hat{\beta}_3 = 1.52243$.

   Calculate the corresponding estimates as in (a) and comment. (3p)

   Solution: Grade four corresponds to category 2 and grade three corresponds to category 1. We have

$$\frac{\pi_2(x)}{\pi_1(x)} = \exp(\alpha_2 + \beta_2 x).$$

   Hence, for students born -97, ($x = 0$), the required estimated ratio is

$$\exp(\hat{\alpha}_2) = \exp(0.37469) \approx 1.4545,$$

   and for students born 98- ($x = 1$), it is

$$\exp(\hat{\alpha}_2 + \hat{\beta}_2) = \exp(0.37469 + 1.57122) \approx 7.0000.$$

   These numbers are very close to the corresponding numbers in (a), indicating a very good model fit. (This is not surprising, since we fit a model with four parameters to a data set with six observations.)

(c) Consider the cumulative logit model

$$\text{logit}\{P(Y \leq j|x)\} = \alpha_j + \beta x,$$

where $j = 1, 2, ..., J-1$, $\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_{J-1}$.
This model was estimated on data (js corresponding to grades as in (a)), resulting in the estimates $\hat{\beta} = -0.6891$, $\hat{\alpha}_1 = -1.1252$, $\hat{\alpha}_2 = 0.9843$.
Interpret the sign on $\hat{\beta}$, calculate the corresponding estimates as in (a) and comment. (3p)

Solution: Observe that the logit function is monotone increasing in its argument. Hence, the negative sign on $\hat{\beta}$ means that given j, the probability of getting a lower grade than what corresponds to j is smaller for students born 98- than for students born -97, i.e. the students born -97 tend to have lower grades than students born 98-. This is also what we see from data.
Now, observe that by inverting the logit function, we have

$$P(Y \leq j|x) = \frac{e^{\alpha_j+\beta x}}{1+e^{\alpha_j+\beta x}} = \frac{1}{1+e^{-\alpha_j-\beta x}},$$

and so, for $j \geq 2$,

$$\begin{aligned} P(Y = j|x) &= P(Y \leq j|x) - P(Y \leq j-1|x) \\ &= \frac{1}{1+e^{-\alpha_j-\beta x}} - \frac{1}{1+e^{-\alpha_{j-1}-\beta x}}, \end{aligned}$$

while

$$P(Y = 1|x) = P(Y \leq 1|x) = \frac{1}{1+e^{-\alpha_1-\beta x}}.$$

This means that the required probability ratios are estimated as, for students born -97, ($x = 0$),

$$\begin{aligned} &\frac{1/\{1+\exp(-\hat{\alpha}_2)\} - 1/\{1+\exp(-\hat{\alpha}_1)\}}{1/\{1+\exp(-\hat{\alpha}_1)\}} \\ =&\frac{1/\{1+\exp(-0.9843)\} - 1/\{1+\exp(1.1252)\}}{1/\{1+\exp(1.1252)\}} \approx 1.97, \end{aligned}$$

and for students born 98- ($x = 1$),

$$\begin{aligned} &\frac{1/\{1+\exp(-\hat{\alpha}_2-\hat{\beta})\} - 1/\{1+\exp(-\hat{\alpha}_1-\hat{\beta})\}}{1/\{1+\exp(-\hat{\alpha}_1-\hat{\beta})\}} \\ =&\frac{1/\{1+\exp(-0.9843+0.6891)\} - 1/\{1+\exp(1.1252+0.6891)\}}{1/\{1+\exp(1.1252+0.6891)\}} \\ \approx& 3.09. \end{aligned}$$

These numbers are much farther away from the estimates in (a) than were the corresponding numbers of (b). It is not surprising that they fit less well, given that the model here has fewer parameters. Also, it seems that the model assumptions are not so well met. The differences between students born -97 or 98- are not of the same type for all grades.

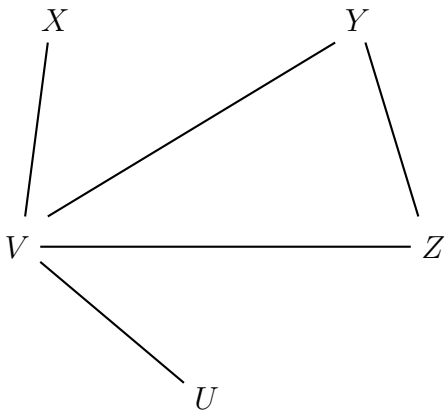

Figure 1: Model, problem 5.

5. Suppose we have variables X, Y, Z, U, V and a loglinear model described by the graph in figure 1.

   (a) Give the name (symbol) of a model that may be described by this graph expressed in a form like (X, Y, YZ) (which is not the model under question here). Also, write down the model equation (in a form like $\log \mu_{ijklm} = \lambda + \lambda_i^X + ...$). (2p)

   Solution: The edges represent interactions and the nodes represent variables. Hence, this is a model with main effects, and pairwise interactions between (X, V), (Y, V), (Y, Z), (Z, V) and (U, V) There might also be a (Y, Z, V) interaction present. (The models with or without it are represented by the same graph.)

   Hence, the model symbol is either (XV, UV, YZ, YV, ZV) or (XV, UV, YZV).

   The model equation for (XV, UV, YZ, YV, ZV) is

$$\begin{aligned}\log \mu_{ijklm} = \lambda &+ \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^U + \lambda_m^V \\ &+ \lambda_{im}^{XV} + \lambda_{lm}^{UV} + \lambda_{jk}^{YZ} + \lambda_{jm}^{YV} + \lambda_{km}^{ZV},\end{aligned}$$

   where μ_{ijklm} is the expected count. For the model (XV, UV, YZV), we add the term λ_{jkm}^{YZV} on the right-hand side.

   (b) Is X independent of Z? (2p)

   Solution: No, because there is a connection between X and Z in the graph via V (and Y). This can also be seen from the model equation, by the presence of the terms λ_{im}^{XV} and λ_{km}^{ZV}.

(c) Is X conditionally independent of Z given V? (2p)

Solution: Yes, because crossing out V, we delete the connection between X and Z in the graph.
(In the model equation, λ_m^V, λ_{im}^{XV}, λ_{km}^{ZV} and λ_{lm}^{UV} are crossed out.)

(d) Is X conditionally independent of Z given Y? (2p)

Solution: No, because crossing out Y, the connection over V still remains.

6. Aggregate data on applicants to graduate school at Berkeley for the six largest departments (d) in 1973 classified by admission (a) and gender (g) was analyzed. The admission variable is 1 for admitted and 0 for rejected. The gender variable is 1 for male and 0 for female.

   Let μ_{ijk} be the expected count in cell (i, j, k), $i, j = 1, 2$, $k = 1, 2, ..., 6$, where i corresponds to admission, j corresponds to gender and k corresponds to department.

   A log linear Poisson model was fit to this data, including all main effects and all two-way interactions. The coefficients with at least one index equal to one were set to zero. The R print from the model estimation is given below.

```
> m=glm(n~a+g+factor(d)+a:g+a:factor(d)+g:factor(d));summary(m)

Call:
glm(formula = n ~ a + g + factor(d) + a:g + a:factor(d) + g:factor(d))

Deviance Residuals:
     1       2       3       4       5       6       7       8       9
 14.50  -14.50  -14.50   14.50   16.50  -16.50  -16.50   16.50    8.25
    10      11      12      13      14      15      16      17      18
 -8.25   -8.25    8.25  -24.75   24.75   24.75  -24.75   12.25  -12.25
    19      20      21      22      23      24
-12.25   12.25  -26.75   26.75   26.75  -26.75

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    327.50      35.84   9.137 0.000263 ***
a              170.00      43.51   3.907 0.011332 *
g             -323.00      43.51  -7.423 0.000699 ***
factor(d)2    -104.00      49.34  -2.108 0.088870 .
factor(d)3    -114.25      49.34  -2.316 0.068433 .
factor(d)4     -73.25      49.34  -1.485 0.197781
factor(d)5    -177.25      49.34  -3.592 0.015669 *
factor(d)6      -3.25      49.34  -0.066 0.950035
a:g            -71.00      32.89  -2.158 0.083349 .
a:factor(d)2   -57.00      56.97  -1.000 0.363015
a:factor(d)3  -271.50      56.97  -4.765 0.005035 **
a:factor(d)4  -261.50      56.97  -4.590 0.005894 **
a:factor(d)5  -279.50      56.97  -4.906 0.004452 **
a:factor(d)6  -445.50      56.97  -7.819 0.000549 ***
g:factor(d)2    91.00      56.97   1.597 0.171105
g:factor(d)3   492.50      56.97   8.644 0.000342 ***
g:factor(d)4   337.50      56.97   5.924 0.001955 **
g:factor(d)5   459.50      56.97   8.065 0.000475 ***
g:factor(d)6   342.50      56.97   6.012 0.001830 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1623)

    Null deviance: 451210  on 23  degrees of freedom
Residual deviance:   8115  on  5  degrees of freedom
AIC: 247.87

Number of Fisher Scoring iterations: 2
```

(a) Explain why the number of degrees of freedom for residual deviance is 5. (1p)

Solution: There are $2 * 2 * 6 = 24$ observed counts. The number of estimated parameters is

$$\begin{aligned}&1 + (2-1) + (2-1) + (6-1)\\&+ (2-1)(2-1) + (2-1)(6-1) + (2-1)(6-1)\\&= 19.\end{aligned}$$

Hence, the number of degrees of freedom is $24 - 19 = 5$.

(b) Test the model vs the saturated model and interpret the result. (2p)

Solution: The saturated model (which is the model above plus the three way interaction) has residual deviance 0 and 0 degrees of freedom. Hence, testing the above model vs the saturated model is performed by comparing the residual deviance above with the χ^2 distribution with 5 degrees of freedom. We have

$$8115 > 20.52 = \chi^2_{0.001}(5),$$

which means that we reject the model at level 0.1%.

The interpretation is that there is room for improving the model. (But this does not seem possible within the loglinear framework with nominal explanatory variables.)

(c) Which logit model for the probability of admittance does this loglinear model correspond to? Based on the R estimation of the loglinear model above, which are the estimated paramaters of this logit model? What is the corresponding number of degrees of freedom? Explain! (4p)

Solution: The loglinear model above may be written as

$$\log \mu_{ijk} = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{ij} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{jk},$$

where X corresponds to admittance, Y corresponds to gender and Z corresponds to department.

It follows that

$$\begin{aligned}&\log\left\{\frac{P(X=1|Y=j,Z=k)}{P(X=0|Y=j,Z=k)}\right\}\\=&\log\left(\frac{\mu_{2jk}}{\mu_{1jk}}\right) = \log\mu_{2jk} - \log\mu_{1jk}\\=&\left(\lambda + \lambda^X_2 + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{2j} + \lambda^{XZ}_{2k} + \lambda^{YZ}_{jk}\right)\\&-\left(\lambda + \lambda^X_1 + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{1j} + \lambda^{XZ}_{1k} + \lambda^{YZ}_{jk}\right)\\=&(\lambda^X_2 - \lambda^X_1) + (\lambda^{XY}_{2j} - \lambda^{XY}_{1j}) + (\lambda^{YZ}_{2k} - \lambda^{YZ}_{1k})\\=&\alpha + \beta^Y_j + \beta^Z_k.\end{aligned}$$

Hence, from the R output (observe that the reference category is obtained if some index equals 1)

$$\begin{aligned}\hat{\alpha} &= \hat{\lambda}_2^X - \hat{\lambda}_1^X = 170.00 - 0 = 170.00,\\ \hat{\beta}_2^Y &= \hat{\lambda}_{22}^{XY} - \hat{\lambda}_{12}^{XY} = -323.00 - 0 = -323.00,\\ \hat{\beta}_2^Z &= \hat{\lambda}_{22}^{YZ} - \hat{\lambda}_{12}^{YZ} = -57.00 - 0 = -57.00,\end{aligned}$$

and similarly, $\hat{\beta}_3^Z = -271.50$, $\hat{\beta}_4^Z = -261.50$, $\hat{\beta}_5^Z = -279.50$ and $\hat{\beta}_6^Z = -445.50$.

The number of degrees of freedom is still 5, because for the logit model, we only have $2 * 6 = 12$ observation cells and the number of parameters is 7, giving $12 - 7 = 5$.

APPENDIX B

Chi-Squared Distribution Values

	Right-Tailed Probability						
df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.80	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.8	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.5

Categorical Data Analysis, Third Edition. Alan Agresti.