NTHU STAT 5230 Final Examination June 17, 2024

Instructions: Attempt all questions. Short and specific answers are preferred. Given explanation when required, but keep it as short and simple as possible. Give only one answer to each question – if you give alternative answers, the worst answer will be graded.

Question A.

Data were collected on 118 female psychiatric patients. The following variables were recorded:

age:	in years
iq:	IQ score
anxiety:	anxiety level (1=none, 2=mild, 3=moderate, 4=severe)
depress:	depression level (1=none, 2=mild, 3=moderate or severe)
sleep:	sleeping normally (1=yes, 2=no)
sex:	lost interest in sex (1=no, 2=yes)
life:	considered suicide (1=no, 2=yes)

Some values in some cases were missing.

(1) (3.5 pts) Consider each variable above as a possible response. For each variable, name a suitable GLM and the distribution assigned to the response when using that variable as the response.

Using just the data on `sleep`, `sex`, and `life`, a 3-way contingency table was constructed:

	life	1	2
sleep	sex		
1	1	0	8
	2	1	5
2	1	55	30
	2	4	7

and a log-linear model was fitted. The following sequential analysis of deviance (i.e., sequentially adding effects) table was obtained:

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			7	155.948	
sleep	1	68.635	6	87.313	< 2.2e-16
life	1	0.910	5	??	0.340022
sex	1	57.779	4	??	2.932e-14
sleep:life	1	16.380	3	??	5.184e-05
sleep:sex	1	7.241	2	??	0.007126
life:sex	1	1.849	1	??	0.173877
sleep:life:sex	1	??	0	????	0.075756

(2) (1 pt) How many cases, at least, have missing values? Explain.

(3) (2 pts) What kind of sampling scheme corresponds to the log-linear model mentioned above (you can refer to schemes 1, 2, 3, and 4 in the lecture note to answer this question)? Please describe a scenario detailing how this sampling scheme can be used to collect these 118 cases.

(4) (1 pt) Provide the value marked as ???? in the analysis of deviance table, and explain why it is this value.

(5) (2 pts) Assuming the p-values in the analysis of deviance table would not change when the order of effects entering the model was changed, describe the nature of the dependence between the 3 variables *in words that a client with no statistical background can easily grasp*. Explain your reasoning.

(6) (2 pts) In this data, would you expect variables `life` and `sex` to be independent? Please use the 3-way contingency table to explain your reasoning in detail.

(7) (1 pt) In the analysis of deviance table, the contribution of the main effect of `life` is not statistically significant. What does this tell us about the distribution of this variable?

(8) (4 pts) When we are particularly interested in the variable `life`, our analysis can be conditioned on the 4 marginal totals with (`sleep`, `sex`) being fixed at (1, 1), (1, 2), (2, 1), and (2, 2), respectively. What type of GLM is suitable in this scenario? Describe the model's response and predictors. Based on your answer to the question (5), what model (i.e., including what effects) would be a good fit in this scenario?

(9) (2 pts) The Mantel-Haenszel (MH) test can be applied to a $2\times2\times k$ contingency table, where the first two variables, denoted as X1 and X2, each have 2 categories, and the third variable, X3, can have k ($k\geq2$) categories. If three MH tests are conducted with X3 being `sleep`, `life`, `sex` respectively, which of the 3 tests will be the least significant (i.e., the most insignificant)? Explain.

(10) (2 pts) For this 3-way contingency table, identify a key benefit of applying the MH tests over the analysis of deviance table to investigate the dependence among the 3 variables, and justify your answer.

An ordinal-by-ordinal log-linear model was fit to the 2-way contingency table constructed using only the data on `anxiety` and `depress`, and the following results were obtained, in which `Oanx` and `Odep` were obtained by the commands `unclass(anxiety)` and `unclass(depression)`, respectively:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0323	0.4442	0.073	0.941976
anxiety2	-0.9087	0.6130	-1.482	0.138238
anxiety3	-5.1643	1.3440	-3.843	0.000122
anxiety4	-12.0088	2.46636	-4.869	1.12e-06
depression2	-2.8626	0.78737	-3.636	0.000277
depression3	-9.1366	1.8804	-4.859	1.18e-06
Oanx:Odep	1.8514	0.3845	4.82	1.5e-06

```
---
    Null deviance: 153.3122  on 11  degrees of freedom
Residual deviance:   3.3504  on  5  degrees of freedom
```

(11) (2 pts) In this analysis, what scores were assigned on the variables `anxiety` and `depress`? Based on the information available, evaluate and comment on whether this score assignment is appropriate.

(12) (1 pt) Is there any evidence that higher levels of anxiety are associated with higher levels of depression? Explain.

(13) (1.5 pts) Evaluate the fitted odds ratio of the 2-by-2 table formed by `anxiety` being 1 and 4 and `depress` being 1 and 3.

A proportional odds logistic regression model was fit to the data using `depress` as a response and predictors as indicated below giving the following output:

```
Coefficients:
        Value      Std. Error   t value
age     -0.01620   0.04648      -0.349
sex     -1.36929   0.66573      -2.057
iq      0.00474    0.00453      1.046
sleep   4.22688    1.09627      3.856

Intercepts:
      Value    Std. Error   t value
1|2   4.784    2.782        1.720
2|3   8.384    2.805        2.989

Residual Deviance: 153.10
```

In this analysis, the original values of predictors from the dataset were utilized (i.e., `sex`=1 or 2 and `sleep`=1 or 2).

(14) (2 pts) Are patients that sleep normally more or less likely to suffer from depression? Explain.

(15) (2 pts) Considering the latent variable viewpoint, what assumptions have been imposed on the latent variable behind `depress` in this model?

A proportional odds logistic regression model using only `sleep` was fit to the data producing the following output:

```
Coefficients:
         Value      Std. Error   t value
sleep    4.36215    1.07977      4.040

Intercepts:
      Value    Std. Error   t value
1|2   6.930    2.096        3.305
2|3   10.349   2.163        4.784

Residual Deviance: 158.29
```

(16) (3 pts) Using this model, estimate the probability that patient who is not sleeping normally, is also not depressed.

(17) (2 pts) Is the model using only `sleep` to be preferred to the model using the four predictors? Explain.

~~Question B.~~

~~Data are generated from the exponential distribution with density $f(y|\lambda) = (1/\lambda)\exp(-(1/\lambda)y)$, where $\lambda>0$ and $y>0$. The exponential distribution is a member of the exponential family which takes the general form:~~

$$f(y|\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right].$$

~~(18) (3 pts) Identify the specific form of θ, ϕ, $a()$, $b()$, and $c()$ for the exponential distribution.~~

~~(19) (2 pts) What is the canonical link and variance function for a GLM with a response following the exponential distribution?~~

~~(20) (1 pt) Identify a practical difficulty that may arise when using the canonical link in this instance [**Hint**: $\lambda>0$].~~