HWA3, Analysis of Categorical Data

Shaobo Jin

1. (1pt) Suppose that we have data that forms an $I \times J \times K \times M$ table for the variables X, Y, Z, and W. We want to build a baseline category model where Y is the response variable. Present a baseline category model that satisfies $X \perp Y \mid (Z, W)$ but not mutual independence. Explain also why such conditional independence holds.
Solution: One such model can be

$$\log\left[\frac{P(Y=j \mid X=i, Z=k, W=m)}{P(Y=J \mid X=i, Z=k, W=m)}\right] = \alpha_{jk} + \beta_{jm} + \gamma_{jkm},$$

where $\alpha_{Jk} = \beta_{Jm} = \gamma_{Jbm} = 0$ for identification. This model implies that

$$P(Y=j \mid X=i, Z=k, W=m) = \frac{\exp(\alpha_{jk} + \beta_{jm} + \gamma_{jkm})}{1 + \sum_{j=1}^{J-1} \exp(\alpha_{jk} + \beta_{jm} + \gamma_{jkm})}$$

for $j \neq J$. Note that

$$\begin{aligned} P(Y=j \mid Z=k, W=m) &= \sum_i P(Y=j \mid X=i, Z=k, W=m) P(X=i \mid Z=k, W=m) \\ &= \sum_i \frac{\exp(\alpha_{jk} + \beta_{jm} + \gamma_{jkm})}{1 + \sum_{j=1}^{J-1} \exp(\alpha_{jk} + \beta_{jm} + \gamma_{jkm})} P(X=i \mid Z=k, W=m) \\ &= \frac{\exp(\alpha_{jk} + \beta_{jm} + \gamma_{jkm})}{1 + \sum_{j=1}^{J-1} \exp(\alpha_{jk} + \beta_{jm} + \gamma_{jkm})}. \end{aligned}$$

Hence, we have $X \perp Y \mid (Z, W)$.

2. (1pt) Consider an $I \times J \times K$ table. Find the connection between the loglinear model of (XYZ) and the baseline category logit model for multinomial sample.
Solution: Consider the loglinear model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

Let $\pi_{ijk} = P(X=i, Y=j, Z=k)$. Then,

$$\begin{aligned} \log\left(\frac{P(Y=j \mid X=i, Z=k)}{P(Y=J \mid X=i, Z=k)}\right) &= \log\left(\frac{\pi_{ijk}}{\pi_{iJk}}\right) = \log\left(\frac{\mu_{ijk}}{\mu_{iJk}}\right) \\ &= \left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}\right) \\ &\quad - \left(\lambda + \lambda_i^X + \lambda_J^Y + \lambda_k^Z + \lambda_{iJ}^{XY} + \lambda_{ik}^{XZ} + \lambda_{Jk}^{YZ} + \lambda_{iJk}^{XYZ}\right) \\ &= \left(\lambda_j^Y - \lambda_J^Y\right) + \left(\lambda_{ij}^{XY} - \lambda_{iJ}^{XY}\right) + \left(\lambda_{jk}^{YZ} - \lambda_{Jk}^{YZ}\right) + \left(\lambda_{ijk}^{XYZ} - \lambda_{iJk}^{XYZ}\right). \end{aligned}$$

This is equivalent to the baseline category logit model

$$\log\left(\frac{P(Y=j \mid X=i, Z=k)}{P(Y=J \mid X=i, Z=k)}\right) = \alpha_j + \beta_{ij}^X + \beta_{kj}^Z + \beta_{ikj}^{XZ}.$$

3. (5pt) Consider the four-way table for car accidents of two time periods

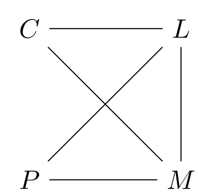
			Period (P)	
Light (L)	Movement (M)	Collision (C)	1	2
daylight	parked	back	712	613
		front	192	179
	moving	back	2257	2373
		front	10749	9768
night/illuminated	parked	back	634	411
		front	95	55
	moving	back	325	283
		front	1256	987
night/dark	parked	back	345	179
		front	46	39
	moving	back	579	494
		front	1018	885

We denote the factors by C, P, M, and L, respectively. We would like to fit a hierarchical loglinear model with generating class (CML, PML).

(a) Write down the equation of the corresponding loglinear model and draw its conditional independence graph.
Solution: The loglinear model is

$$\log \mu_{ijkl} = \lambda + \lambda_i^C + \lambda_j^P + \lambda_k^M + \lambda_l^L + \lambda_{ik}^{CM} + \lambda_{il}^{CL} + \lambda_{kl}^{ML} + \lambda_{jk}^{PM} + \lambda_{jl}^{PL} + \lambda_{ikl}^{CML} + \lambda_{jkl}^{PML}.$$

The CIG is



(b) Is the model a graphical model?
Solution: The maxcliques are (CML, PML), which is the same as the generating class. Hence, it is a graphical model.

(c) Find a minimal sufficient of the model.
Solution: Since the model is a graphical model, $\{n_{i+kl}\}$ and $\{n_{+jkl}\}$ are minimal sufficient.

(d) Find the multigraph of the model.
Solution: The multigraph is

$$CML =\!=\!=\!=\!= PML$$

(e) Is the model decomposable or nondecomposable?
Solution: We can find a cycle $C - M - P - L - C$ with the chord $L - M$. Hence, the graph is chordal, which implies that it is decomposable.

(f) Can you find the closed form expression of the MLE of μ_{ijkl}? If so, derive such expression. Otherwise, state the reason.
Solution: Since the model is decomposable, we can find the explicit expression of joint probabilities as

$$\pi_{ijkl} = \frac{\pi_{i+kl}\pi_{+jkl}}{\pi_{++kl}}.$$

Hence,

$$\mu_{ijkl} = n\pi_{ijkl} = n\frac{\pi_{i+kl}\pi_{+jkl}}{\pi_{++kl}} = \frac{\mu_{i+kl}\mu_{+jkl}}{\mu_{++kl}}.$$

The MLE of μ_{ijkl} is $\hat{\mu}_{ijkl} = n_{i+kl}n_{+jkl}/n_{++kl}$.

(g) State also the which variable(s) you need to conditional on in order for P and C to be independent.
Solution: Note that $\{L, M\}$ separates C and P. Hence, $C \perp P \mid \{L, M\}$.

(h) Fit the hierarchical loglinear model with generating class (PML, CML) and test the assumption of conditional independence of P and C.
Solution:

```
## Code our data
Data <- data.frame(count = c(712, 192, 2257, 10749, 613, 179, 2373, 9768,
                             634, 95, 325, 1256, 411, 55, 283, 987,
                             345, 46, 579, 1018, 179, 39, 494, 885),
                   collision = rep(c("back", "front"), times = 12),
                   period = rep(rep(c("1", "2"), each = 4), times = 3),
                   move = rep(rep(c("parked", "moving"), each = 2), times = 6),
                   light = rep(c("daylight", "nightilluminated", "nightdark"), each = 8))
## Fit model using GLM
PoiReg <- glm(count ~ collision + period + move + light +
                      collision:move + collision:light + move:light +
                      period:move + period:light +
                      collision:move:light + period:move:light,
                      data = Data, family=poisson())
PoiReg

##
## Call:  glm(formula = count ~ collision + period + move + light + collision:move +
##     collision:light + move:light + period:move + period:light +
##     collision:move:light + period:move:light, family = poisson(),
##     data = Data)
##
## Coefficients:
##                                    (Intercept)
##                                        7.78098
##                                 collisionfront
##                                        1.48870
##                                        period2
##                                       -0.06882
##                                     moveparked
##                                       -1.22102
##                                 lightnightdark
##                                       -1.42522
##                         lightnightilluminated
##                                       -1.96042
##                     collisionfront:moveparked
##                                       -2.76166
##                 collisionfront:lightnightdark
##                                       -0.91572
##         collisionfront:lightnightilluminated
##                                       -0.18330
##                     moveparked:lightnightdark
##                                        0.68364
##             moveparked:lightnightilluminated
##                                        1.85800
##                            period2:moveparked
```

```
##                                          -0.06345
##                             period2:lightnightdark
##                                          -0.07795
##                     period2:lightnightilluminated
##                                          -0.15022
##        collisionfront:moveparked:lightnightdark
##                                           0.36985
## collisionfront:moveparked:lightnightilluminated
##                                          -0.48487
##                 period2:moveparked:lightnightdark
##                                          -0.37400
##        period2:moveparked:lightnightilluminated
##                                          -0.16500
##
## Degrees of Freedom: 23 Total (i.e. Null);  6 Residual
## Null Deviance:      69680
## Residual Deviance: 26.49  AIC: 255.3
```

We can also fit the model using loglm().

```
## We can also fit the model using loglm
library(MASS)
logLin <- loglm(count ~ collision + period + move + light +
                      collision:move + collision:light + move:light +
                      period:move + period:light +
                      collision:move:light + period:move:light,
                      data = Data, fitted = TRUE, param = TRUE)
logLin

## Call:
## loglm(formula = count ~ collision + period + move + light + collision:move +
##     collision:light + move:light + period:move + period:light +
##     collision:move:light + period:move:light, data = Data, fitted = TRUE,
##     param = TRUE)
##
## Statistics:
##                       X^2 df     P(> X^2)
## Likelihood Ratio 26.48513  6 0.0001807631
## Pearson          26.61533  6 0.0001709145
```

Test the conditional independence of C and P can be viewed as test whether the model fits the data well. Note that

```
deviance(PoiReg)

## [1] 26.48513

qchisq(0.95, PoiReg$df.residual)

## [1] 12.59159
```

The model does not fit the data as well as the saturated model. Hence, we may not favor the conditional independence assumption.

(i) From the model that fitted above, compute the conditional PM odds ratio given *front* and *night/illuminated* without using the predict() function. You can compare your results with the odds ratio from the predict() function.

Solution: The conditional odds ratio can be computed by two methods. If we fit the loglinear model in glm(), we can use the coefficients to compute the conditional log odds ratio as follows.

```
# period "1" is the reference category.
# move "moving" is the reference category.
park.p1 <- 0  # parked, period 1
move.p2 <- 0 # moving, period 2
park.p2 <- (coef(PoiReg)[["period2:moveparked"]] +
            coef(PoiReg)[["period2:moveparked:lightnightilluminated"]]) # parked, period 2
move.p1 <- 0 # moving, period 1
park.p1 + move.p2 - park.p2 - move.p1

## [1] 0.2284474
```

If we use the loglm() function, then the log odds ratio can be computed as follows.

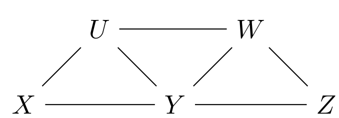
```
park.p1 <- coef(logLin)[["period.move"]]["1", "parked"] +
    coef(logLin)[["period.move.light"]]["1", "parked" ,"nightilluminated"]
move.p2 <- coef(logLin)[["period.move"]]["2", "moving"] +
    coef(logLin)[["period.move.light"]]["2", "moving" ,"nightilluminated"]
park.p2 <- coef(logLin)[["period.move"]]["2", "parked"] +
    coef(logLin)[["period.move.light"]]["2", "parked" ,"nightilluminated"]
move.p1 <- coef(logLin)[["period.move"]]["1", "moving"] +
    coef(logLin)[["period.move.light"]]["1", "moving" ,"nightilluminated"]
park.p1 + move.p2 - park.p2 - move.p1

## [1] 0.2284474
```

The conditional odds ratio is then $\exp(0.2284474)$.

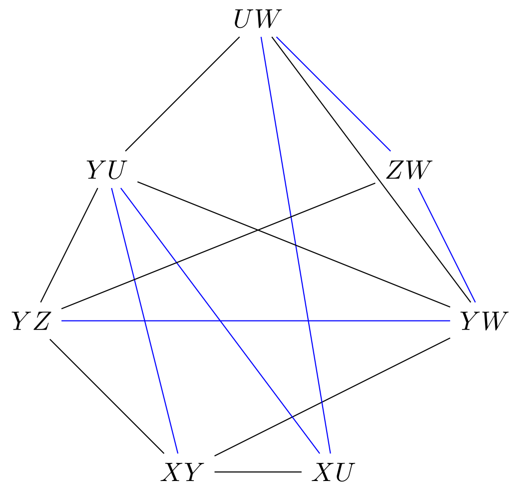
4. (3pt) Consider the hierarchical loglinear model with generating class $(XY, XU, YZ, YW, YU, ZW, UW)$.

   (a) Suppose that the odds ratio $\theta_{is(jkl)}$ for the XU association in the partial table $Y = j$, $Z = k$, and $W = l$ is 2. Can you tell what is the XU association in the partial table $Y = j$, and $W = l$?
   Solution:



   Hence, $Z \perp (X, U) \mid (Y, W)$. After collapsing Z, the association in (X, U) is unchanged. Hence, $\theta_{is(jl)} = 2$.

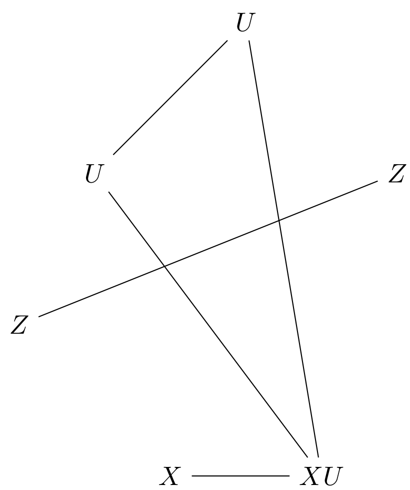
   (b) Find a maximum spanning tree of its multigraph.
   Solution: The multigraph is



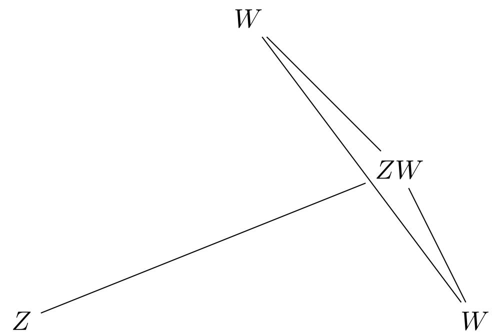
A maximum spanning tree is $[XY]\,[YU]\,[XU]\,[UW]\,[ZW]\,[YW]\,[YZ]$. In this case, the maximum spanning tree is not unique.

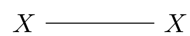
(c) Use the multigraph to determine whether the model decomposable.
Solution: Number of indices added over vertices is 14. Number of indices added over branches is 6. Number of factors is 5. Hence, the model is not decomposable.

(d) Find a fundamental conditional independence set.
Solution: If we remove $\{Y, W\}$, then we obtain



We obtain two disconnected components. Hence, $[X, U \otimes Z \mid Y, W]$. If we remove $\{Y, U\}$, then we obtain





We still obtain two disconnected components. Then, $[X \otimes Z, W \mid Y, U]$.