Categorical Data Analysis Note

Edgard Maboudou

Chapter 7. Log Linear Model

– All variables are categorical (response and explanatory).
– Analyzing multi-way tables can be complicated as we have p nominal variables $X_1, X_2, \ldots, X_p$ where ith variable X_i has I_i levels.
– So we have $I_1 \times I_2 \cdots \times I_p$ cells in our table.
– Are the variables related?
– How are they related?
– Are there interactions? A model with no interaction is simpler.
– For example, with $p = 2$, we have $I \times J$ tables and n observations.
– If X and Y are independent,

$$\pi_{ij} = \pi_{i+}\pi_{+j}.$$

– In terms of expected frequencies

$$\mu_{ij} = E[n_{ij}] = n\pi_{ij} = n\pi_{i+}\pi_{+j}$$

so

$$\log(\mu_{ij}) = \log(n) + \log(\pi_{i+}) + \log(\pi_{+j}) = \lambda + \lambda_i^X + \lambda_j^Y,$$

where

- λ is overall effect,
- λ_i^X is effect due to ith level of X (row effect),
- λ_j^X is effect due to jth level of Y (column effect).

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y,$$

– X and Y are independent implies this loglinear model holds.
– Since $\sum_i \pi_{i+} = 1$, one λ_i^X is redundant, so you only need $I - 1$ of them.
– In the same way, one λ_j^Y is redundant, so you only need $J - 1$ of them.
– In R, set constraint and get solution for that constraint.
– i.e. the solutions for λ_i^X's and λ_j^Y's are not unique. So we usually solve them with some constraint.

Example: Data on Relationship between College Enrollment and Political Affiliation

```
> #Create contingency table
> n.table<-array(c(34,31,19,23,61,19,23,39,16,17,16,12), dim=c(4,3),
+ dimnames=list(enroll = c("art&sci", "eng","agric", "educ"),
+ pol_aff = c("rep", "dem","indep")))
> n.table
         pol_aff
enroll     rep  dem indep
  art&sci  34   61    16  (111)
  eng      31   19    17  (67)
  agric    19   23    16  (58)
  educ     23   39    12  (74)
         (107) (142) (61)
```

– The code used to fit the loglinear model sets first level to 0 (default for R).

```
>  #Data needs to be in a data.frame format for glm()
>  counts<-c(34,31,19,23,61,19,23,39,16,17,16,12)
>  enroll<-rep(c("art&sci", "eng", "agric", "educ"),3)
>  pol_aff<-c(rep("rep",4),rep("dem",4), rep("indep",4))
>  data<-data.frame(enroll, pol_aff, counts)
>  data
    enroll pol_aff counts
1  art&sci     rep     34
2      eng     rep     31
3    agric     rep     19
4     educ     rep     23
5  art&sci     dem     61
6      eng     dem     19
7    agric     dem     23
8     educ     dem     39
9  art&sci   indep     16
10     eng   indep     17
11   agric   indep     16
12    educ   indep     12
>  #Show how to convert back to a contingency table
>  nn.table<-xtabs(formula = counts ~ enroll + pol_aff,data = data)
>  nn.table
```

```
         pol_aff
enroll    dem indep rep
  agric    23    16  19
  art&sci  61    16  34
  educ     39    12  23
  eng      19    17  31
> #Easier way to get the data in the correct form
> data2<-as.data.frame(as.table(n.table))
> data2
    enroll pol_aff Freq
1  art&sci     rep   34
2      eng     rep   31
3    agric     rep   19
4     educ     rep   23
5  art&sci     dem   61
6      eng     dem   19
7    agric     dem   23
8     educ     dem   39
9  art&sci   indep   16
10     eng   indep   17
11   agric   indep   16
12    educ   indep   12
> mod.fit<-glm(formula = counts ~ enroll + pol_aff, data = data2, family = poiss
+     na.action = na.exclude, control = list(epsilon =0.0001, maxit = 50, trace
Deviance = 16.78994 Iterations - 1
Deviance = 16.39044 Iterations - 2
Deviance = 16.39008 Iterations - 3
> summary(mod.fit)

Call:
glm(formula = counts ~ enroll + pol_aff, family = poisson(link = log),
    data = data2, na.action = na.exclude, control = list(epsilon = 1e-04,
        maxit = 50, trace = T))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2714  -0.7091  -0.3707   1.0740   1.5556

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.6458     0.1230  29.643  < 2e-16 ***
```

```
enrolleng     -0.5048     0.1546  -3.266 0.001093 **
enrollagric   -0.6491     0.1620  -4.007 6.16e-05 ***
enrolleduc    -0.4055     0.1501  -2.702 0.006894 **
pol_affdem     0.2830     0.1280   2.211 0.027022 *
pol_affindep  -0.5620     0.1604  -3.504 0.000459 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 69.607  on 11  degrees of freedom
Residual deviance: 16.390  on  6  degrees of freedom
AIC: 88.296

Number of Fisher Scoring iterations: 3
```

– From the output, the estimated model is

$$\log(\hat{\mu}_{ij}) = \hat{\lambda} + \hat{\lambda}_i^C + \hat{\lambda}_j^P = 3.6458 + \hat{\lambda}_i^C + \hat{\lambda}_j^P,$$

where $\hat{\lambda}_{AS}^C = 0$, $\hat{\lambda}_E^C = -0.5048$, $\hat{\lambda}_{Ag}^C = -0.6491$, $\hat{\lambda}_{Ed}^C = -0.4055$, and $\hat{\lambda}_r^P = 0$, $\hat{\lambda}_d^P = 0.2830$, $\hat{\lambda}_i^P = -0.5620$.
– This model is similar to a regression formulation

$$\log(\hat{\mu}) = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 z_1 + \hat{\beta}_5 z_2$$

– To find the estimates for μ_i, use the “predict” function the same way as in Ch 4.
– For example,

$$\log(\hat{\mu}_{11}) = \hat{\lambda} + \hat{\lambda}_1^C + \hat{\lambda}_1^P = 3.6458 + 0 + 0 \quad \mu_{11} = e^{3.6458} = 38.313.$$

$$\log(\hat{\mu}_{21}) = \hat{\lambda} + \hat{\lambda}_1^C + \hat{\lambda}_1^P = 3.6458 - 0.5048 + 0, \quad \mu_{21} = e^{3.141} = 23.126.$$

– Be careful with the order that R writes the prediction out using “predict”.
– Also, one can find the Pearson and standardized Person residuals as described before.
– This is done in R as

```
> #mod.fit$fitted.values provides the same values
> save.predict<-predict(object = mod.fit,type="response")
> save.pearson<-residuals(object = mod.fit,type="pearson")
> h<-lm.influence(model = mod.fit)$h
> standard.pearson<-save.pearson/sqrt(1-h)
> save.all<-data.frame(data2, predict = round(save.predict,4), pearson =
+     round(save.pearson,4), standard.pearson = round(standard.pearson,4))
> save.all
    enroll pol_aff Freq predict pearson standard.pearson
1  art&sci     rep   34 38.3129 -0.6968          -1.0745
2      eng     rep   31 23.1259  1.6374           2.2857
3    agric     rep   19 20.0194 -0.2278          -0.3122
4     educ     rep   23 25.5419 -0.5030          -0.7122
5  art&sci     dem   61 50.8452  1.4241           2.4144
6      eng     dem   19 30.6904 -2.1102          -3.2386
7    agric     dem   23 26.5677 -0.6922          -1.0429
8     educ     dem   39 33.8968  0.8765           1.3646
9  art&sci   indep   16 21.8419 -1.2500          -1.7406
10     eng   indep   17 13.1839  1.0510           1.3248
11   agric   indep   16 11.4129  1.3578           1.6803
12    educ   indep   12 14.5613 -0.6712          -0.8583
```

– As the counts for a contingency table are being modeled, it is often of interest to examine the prediction and Pearson residuals in the same contingency table format.

```
> xtabs(predict ~ enroll + pol_aff, data = save.all)
         pol_aff
enroll        rep     dem   indep
  art&sci 38.3129 50.8452 21.8419
  eng     23.1259 30.6904 13.1839
  agric   20.0194 26.5677 11.4129
  educ    25.5419 33.8968 14.5613
> xtabs(pearson ~ enroll + pol_aff, data = save.all)
         pol_aff
enroll        rep     dem   indep
  art&sci -0.6968  1.4241 -1.2500
  eng      1.6374 -2.1102  1.0510
  agric   -0.2278 -0.6922  1.3578
  educ    -0.5030  0.8765 -0.6712
> xtabs(standard.pearson ~ enroll + pol_aff, data = save.all)
         pol_aff
enroll        rep     dem   indep
  art&sci -1.0745  2.4144 -1.7406
  eng       2.2857 -3.2386  1.3248
  agric    -0.3122 -1.0429  1.6803
  educ     -0.7122  1.3646 -0.8583
```

– Make sure to compare the estimate cell counts ($\hat{\mu}_{ij}$) to the observed cell counts (n_{ij}).
– The estimated odds ratio can be found from the $\hat{\mu}_{ij}$'s.
– Measures of overall goodness-of-fit can be found using the Pearson and LRT statistics (as before).

```
> #Calculate goodness-of-fit measures
> pearson.stat<-sum(save.pearson^2)
> cat("X^2 =", round(pearson.stat, 4),"with p-value =", round(1-pchisq(pearson.s
+       mod.fit$df.residual),4), "\n")
X^2 = 16.1613 with p-value = 0.0129
> dev<-mod.fit$deviance
> cat("G^2 =", round(dev, 4),"with p-value =",round(1-pchisq(dev,mod.fit$df.resi
G^2 = 16.3901 with p-value = 0.0118
```

– From the R output above, the P-values are small. This indicates that the loglinear model under independence does not fit the data well.
– Using loglinear independence model, the log odds of Y being at j as opposed to k when $X = i$ is:

$$\log\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \log(\mu_{ij}) - \log(\mu_{ik}) = \lambda_j^Y - \lambda_k^Y,$$

This does not depend on the i.
– That is, this loglinear model is equivalent to the logit model

$$\log\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \alpha.$$

– We get the odds by exponentiating

$$\frac{\mu_{ij}}{\mu_{ik}} = e^{\lambda_j^Y - \lambda_k^Y}.$$

– Back to the example, let's use our model to estimate the odds of being a democrat as opposed to independent when in the college of education.

$$\frac{\pi_{Ed,d}}{\pi_{Ed,i}} = e^{\hat{\lambda}_d^P - \hat{\lambda}_i^P} = e^{0.2830-(-0.5620)} = e^{0.845} = 2.328.$$

– Assume the independent model,

```
           pol_aff
enroll      rep   dem indep
  educ       23    39   12  (74)
           (107) (142) (61) (310)
```

$$\frac{\mu_{Ed,d}}{\mu_{Ed,i}} = \frac{74 \times 142}{310} \cdot \frac{310}{74 \times 61} = 2.328.$$

– If the independent model does not hold, then we need a more complicated model, saturated model.

Saturated Model

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$

where λ_{ij}^{XY} is the interaction term.
– If all interaction terms are 0, then they are not needed in the model and the model will reduce to the independence model.
– We have as many parameters as expected cell counts.
– For this reason, this model will always hold and is called the saturated loglinear model.

$$(I-1) \times (J-1) + (I-1) + (J-1) + 1 = I \times J.$$

–Some of the λ_{ij}^{XY} are redundant, so get estimates by imposing constraints such as

$$\sum_i \lambda_{ij}^{XY} = 0 \text{ for each } j \text{ (row sum = 0)}, \ \sum_j \lambda_{ij}^{XY} = 0 \text{ for each } i \text{ (column sum = 0)}.$$

Local Odds Ratio, θ_{ij}

$$\theta_{ij} = \frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}}.$$

– Computed from 2 × 2 table within $I \times J$ table.
– Local when two adjacent rows and two adjacent columns.
– Ex. We can compute 4 local odds ratios in a 3 × 3 table.

– Now

$$\begin{aligned}\log\theta_{ij} &= \log\left(\frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}}\right)\\ &= \\ &= \\ &= \end{aligned}$$

– Then

$$\theta_{ij} = e^{\lambda_{ij}^{XY}+\lambda_{i+1,j+1}^{XY}-\lambda_{i,j+1}^{XY}-\lambda_{i+1,j}^{XY}}.$$

Example fit the saturated model

$$\lambda_{as,d}^{CP}=0 \quad \lambda_{eng,d}^{CP}=-1.07406 \quad \lambda_{ag,d}^{CP}=-0.39346 \quad \lambda_{edu,d}^{CP)}=-0.05645$$

$$\lambda_{as,i}^{CP}=0 \quad \lambda_{eng,i}^{CP}=0.153 \quad \lambda_{ag,i}^{CP}=0.58192 \quad \lambda_{edu,i}^{CP}=0.10318$$

$$\lambda_{as,r}^{CP}=0 \quad \lambda_{eng,r}^{CP}=0 \quad \lambda_{ag,r}^{CP}=0 \quad \lambda_{edu,r}^{CP}=0$$

– Estimated odds ratio comparing ag and ed, d and i.

```
         pol_aff
enroll     rep  dem    indep
  art&sci  34   61      16  (111)
  eng      31   19      17  (67)
  agric    19   *23    *16  (58)
  educ     23   *39    *12  (74)
          (107) (142)   (61)
```

$\theta_{32} = e^{\lambda_{ag,d}^{CP}+\lambda_{edu,i}^{CP}-\lambda_{ag,i}^{CP}-\lambda_{edu,d}^{CP}} =$

– Since the model fits the data perfectly, this is the same odds ratio we get by using the observed frequencies.

– For two-way tables, we basically have 2 possible loglinear models.

- Independence model (simple)
- Saturated model (most complicated)

– When we have more than 2 variables, there will be unsaturated models besides the independence model.
– If these models fit, they simplify the interpretation of association.

Models for Three-Way Tables

– X, Y, Z in a $I \times J \times K$ table.
– We consider hierarchical models: include all lower order terms on variables in any higher order terms, i.e.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2$$

Even if $\beta_1 = \beta_3 = 0$, you need to keep β_1 because you have the interaction term β_4 with $x_1 x_2$.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_1 x_2$$

– Possible models (Total of 9)

1. Mutual Independence Model.

   $$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

   has main effect terms but no interaction terms. If this model fits, the variables are not related to one another. This model is denoted as (X, Y, Z).

2. (XY, Z) Model.

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

   has interaction term λ_{ij}^{XY} implies X and Y are dependent given Z. No λ_{ik}^{XZ} term implies X and Z are independent given Y. No λ_{jk}^{YZ} term implies Y and Z are independent given X.

   $$\theta_{i(j)k} = \theta_{(i)jk} = 1.$$

   this model is denoted as (XY, Z), (XZ, Y), (X, YZ). The odds ratio does not depend on k:

   $$\theta_{ij(k)} = e^{\lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}}.$$

3. (XY, XZ) Model.

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

   has two two-way interaction terms: X and Y are dependent given Z. X and Z are dependent given Y. Y and Z are independent given X, so $\theta_{(i)jk} = 1$.

$$\theta_{i(j)k} = e^{\lambda_{ik}^{XZ} + \lambda_{i+1,k+1}^{XZ} - \lambda_{i,k+1}^{XZ} - \lambda_{i+1,k}^{XZ}}.$$

$$\theta_{ij(k)} = e^{\lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}}.$$

   this model is denoted (XY, XZ), (XY, YZ), or (XZ, YZ).

4. Pairwise Dependent Model (XY, XZ, YZ). This is the only one Model with all two-way interactions.

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

   It is denoted (XY, XZ, YZ). The odds ratio is

$$\theta_{(i)jk} = e^{\lambda_{jk}^{YZ} + \lambda_{j+1,k+1}^{YZ} - \lambda_{j,k+1}^{YZ} - \lambda_{j+1,k}^{YZ}}.$$

   When it does not depend on i, we can say that

$$\theta_{(1)jk} = \theta_{(2)jk} \cdots = \theta_{(I)jk}.$$

   Similarly,

$$\theta_{i(1)k} = \theta_{i(2)k} \cdots = \theta_{i(J)k}.$$

$$\theta_{ij(1)} = \theta_{ij(2)} \cdots = \theta_{ij(K)}.$$

   We have homogenous association. Hence, simplifies the interpretation. For the 3rd equation, the relationship between X and Y given Z does not depend on what level Z is fixed at.

5. Saturated/Complete Model (XYZ). (as many parameters as cells)

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

   It fits data perfectly (if this is the only model that fits, then our modeling efforts have been unsuccessful). The data is too complex to fit and model data does not exhibit any simplifying structure.

If we want to test if the λ_{ijk}^{XYZ} is necessary, we compare model (XYZ) with model (XY, XZ, YZ).

Conditional Tests

– Test a reduced model M_2 from a complex model M_1.

$$G^2(M_2|M_1) = G^2(M_2) - G^2(M_1).$$

with $df(M_2|M_1) = df(M_2) - df(M_1)$.

Example: Alcohol, Cigarette, and Marijuana Use

– Let M_1 be model (AC, AM, CM) and M_2 be (AM, CM).
– The complete model M_1 is fitted in R as follow:

```
> table7.3<-data.frame(expand.grid(marijuana=factor(c("yes","no"),levels=c("yes"
+ cigarette=factor(c("yes","no"),levels=c("yes","no")),
+ alcohol=factor(c("yes","no"),levels=c("yes","no"))), count=c(911,538,44,456,3,
> options(contrasts=c("contr.treatment","contr.poly"))
> fit.Ha<-glm(count~.^2,data=table7.3,family=poisson)
> summary(fit.Ha)

Call:
glm(formula = count ~ .^2, family = poisson, data = table7.3)

Deviance Residuals:
       1         2         3         4         5         6         7         8
 0.02044  -0.02658  -0.09256   0.02890  -0.33428   0.09452   0.49134  -0.03690

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              6.81387    0.03313 205.699  < 2e-16 ***
marijuanano             -0.52486    0.05428  -9.669  < 2e-16 ***
cigaretteno             -3.01575    0.15162 -19.891  < 2e-16 ***
alcoholno               -5.52827    0.45221 -12.225  < 2e-16 ***
marijuanano:cigaretteno  2.84789    0.16384  17.382  < 2e-16 ***
marijuanano:alcoholno    2.98601    0.46468   6.426 1.31e-10 ***
cigaretteno:alcoholno    2.05453    0.17406  11.803  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2851.46098  on 7  degrees of freedom
```

```
Residual deviance:    0.37399  on 1  degrees of freedom
AIC: 63.417

Number of Fisher Scoring iterations: 4
```

– Next, we fit the reduced model M_2 by using R

```
> fit.Ho<-glm(count~(alcohol+marijuana)^2+(cigarette+marijuana)^2,data=table7.3,
> summary(fit.Ho)

Call:
glm(formula = count ~ (alcohol + marijuana)^2 + (cigarette +
    marijuana)^2, family = poisson, data = table7.3)

Deviance Residuals:
       1         2         3         4         5         6         7         8
 0.05836   4.57017  -0.26193  -4.34413  -0.86631  -9.77162   2.22872   6.83535

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              6.81261    0.03316 205.450   <2e-16 ***
alcoholno               -5.25227    0.44837 -11.714   <2e-16 ***
marijuanano             -0.72847    0.05538 -13.154   <2e-16 ***
cigaretteno             -2.98919    0.15111 -19.782   <2e-16 ***
alcoholno:marijuanano    4.12509    0.45294   9.107   <2e-16 ***
marijuanano:cigaretteno  3.22431    0.16098  20.029   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2851.46  on 7  degrees of freedom
Residual deviance:  187.75  on 2  degrees of freedom
AIC: 248.8

Number of Fisher Scoring iterations: 5
```

– $G^2(M_2|M_1)$ can also be computed directly in R

```
> G.sq<-fit.Ho$deviance-fit.Ha$deviance
> G.sq.df<-fit.Ho$df.residual- fit.Ha$df.residual
> cat("G^2 =", round(G.sq,4), "with p-value =", round(1-pchisq(G.sq, G.sq.df),4)
G^2 = 187.3803 with p-value = 0
```

– Note that the difference between M_1 and M_2 is that $\lambda^{AC} = 0$ in M_2, which implies A and C are conditionally independent given M.
– So $G^2(M_2|M_1)$ is a test for conditional independence of A and C. This is an alternative to the Cochran-Mantel-Haenszel Test (Ch3).

CI for Odds Ratio

A 95% confidence interval for the log of a conditional odds ratio is

$$\log \hat{\theta} \pm 1.96 \cdot \hat{\sigma}_{\log \hat{\theta}}.$$

CI for θ is obtained by exponentiating.

Ex.

$$\theta_{ij(k)} = e^{\lambda_{ij}^{AC}+\lambda_{i+1,j+1}^{AC}-\lambda_{i,j+1}^{AC}-\lambda_{i+1,j}^{AC}}.$$

From the output,

Interpretation: There is a strong positive association between cigarette use and alcohol use, both for users and non users of marijuana.

Four-Way and Higher Order Table

W, X, Y, Z in an $H \times I \times J \times K$ table.
– Independence model (W, X, Y, Z)

$$\log(\mu_{hijk}) = \lambda + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

Logit Models

– Logit models with nominal explanatory variables are equivalent to particular loglinear models.
– Logit models have one of the variables playing the role of response with the other being explanatory variables.
– Consider 3 variables: X, Y, Z and the loglinear model (XY, XZ)

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

– If Y has 2 levels and is the response, then the logit model equivalent to this loglinear model is

$$\log\left(\frac{\mu_{i1k}}{\mu_{i2k}}\right) = (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) = \alpha + \beta_i^X.$$

– From the out put of log linear, you will get $\alpha = \lambda_1^Y - \lambda_2^Y$, $\beta_1^X =$, and $\beta_1^Z =$. If the best model is (XY, YZ, XZ):

$$\begin{aligned} \log\left(\frac{\mu_{i1k}}{\mu_{i2k}}\right) &= \\ &- \\ &= \\ &= \end{aligned}$$

Example

Example Continued

Modeling Ordinal Associations

– Our loglinear models have treated all variables as nominal.
– If in fact, they are ordinal, the models ignore this information.
– Exploit ordinality of the variables gives better models.
– Two-way tables with nominal variables, we have 2 possible models

1. Independent model: variables aren't related.

2. Saturated model: no simplification possible.

For ordinal variables, there is a model in between these 2 models, i.e. it is possible to have a model where there is a simplified interpretation of the relationship between X and Y.

–Suppose X and Y are ordinal with I levels for X and J levels for Y.
–Assign scores to these levels:
scores for X: $u_1 < u_2 \cdots < u_I$
scores for Y: $v_1 < v_2 \cdots < v_J$
–When not appropriate use the counting numbers:
scores for X: $1 < 2 \cdots < I$
scores for Y: $1 < 2 \cdots < J$

Uniform association model (linear-by-linear association)

– Saturated loglinear model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

– The uniform association model is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j.$$

Only one parameter for interaction,

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta ij.$$

$\beta u_i v_j$ (or βij) represents the deviation of $\log \mu_{ij}$ from the independent model.
– This term is linear in the Y scores at a fixed level of X and linear in the X scores at a fixed level of Y.
– So it is called linear-by-linear association model ($L \times L$).
– If $\beta = 0$, then we get back the independence model.
– For fixed j, a one level change in i (from i to $i+1$) changes the interaction by $\beta \times j$.
– $\beta > 0$ implies Y tends to increase as X increases.
– $\beta < 0$ implies Y tends to decrease as X increases.

Interpretation of the actual β value

Look at log odds ratio for adjacent cells (local odds ratio):

$$\begin{aligned}\log\theta_{ij} &= \log\left(\frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i+1,j}\mu_{i,j+1}}\right)\\ &= \\ &= \\ &= \\ &= \end{aligned}$$

Since β does not depend on i and j, so all the local odds ratios are the same $\theta_{ij} = e^{\beta}$. So we call it uniform.
– A 1 unit increase in level is the same at each level.
– θ_{ij} doesn't depend on i and j.

Example Sex Opinions P230

– Using row scores $\{1, 2, 3, 4\}$ and column scores $\{1, 2, 3, 4\}$, the model fit is:

```
> y<-c(81,68,60,38,24,26,29,14,18,41,74,42,36,57,161,157)
> sex<-c(rep(1,4),rep(2,4),rep(3,4),rep(4,4))
> u<-sex
> sex<-factor(sex)
> birth<-rep(seq(1,4),4)
> v<-birth
> birth<-factor(birth)
> fit<-glm(y~sex+birth+u:v,family=poisson)
> summary(fit)

Call:
glm(formula = y ~ sex + birth + u:v, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.35834  -0.91606   0.07972   0.61648   1.57618

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.10684    0.08951  45.881  < 2e-16 ***
```

```
sex2        -1.64596    0.13473 -12.216  < 2e-16 ***
sex3        -1.77002    0.16464 -10.751  < 2e-16 ***
sex4        -1.75369    0.23432  -7.484 7.20e-14 ***
birth2      -0.46411    0.11952  -3.883 0.000103 ***
birth3      -0.72452    0.16201  -4.472 7.74e-06 ***
birth4      -1.87966    0.24910  -7.546 4.50e-14 ***
u:v          0.28584    0.02824  10.122  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 431.078  on 15  degrees of freedom
Residual deviance:  11.534  on  8  degrees of freedom
AIC: 118.21

Number of Fisher Scoring iterations: 4
```

– The fitted values are:

```
> fit$fitted.values
        1         2         3         4         5         6         7         8
 80.85658  67.65406  69.39574  29.09363  20.75004  23.10650  31.54350  17.59996
        9        10        11        12        13        14        15        16
 24.39370  36.15178  65.68137  48.77315  32.99969  65.08766 157.37940 155.53326
```

Using row scores $\{1, 2, 3, 4\}$ and column scores $\{1, 2, 3, 4\}$, the estimated local odds ratio for the UA model is $e^{0.286} = 1.331$. This is confirmed from R

```
> oddrat<-exp(coef(fit)["u:v"]);oddrat
     u:v
1.330873
```

A Wald CI is obtained from the $SE_{\hat{\beta}} = 0.028$

$$\hat{\beta} \pm 1.96 \cdot SE_{\hat{\beta}} = 0.286 \pm 1.96 \times 0.028 \ .$$

The 95% CI for the odds ratio is obtained by exponentiating the end points of the CI for $\log \theta$. In R, we have

```
> rat<-summary(fit)$coef["u:v",1];rat
[1] 0.2858355
```

```
> se<-summary(fit)$coef["u:v",2];se
[1] 0.02823796
> logci<-rat+1.96*c(-1,1)*se;logci
[1] 0.2304891 0.3411819
> ci<-exp(logci);ci
[1] 1.259216 1.406609
```

According to the estimate, the association is positive, implying that subjects with more favorable attitudes about teenage birth control also tended to have more tolerant attitudes about premarital sex.

But the association is rather weak.

The odds of responding in the adjacent higher category on one variable are 1.33 times more likely if subject responded in the higher category of the other variable (vs the lower category).

However, with distance in categories, the odds ratios are larger. For example, the odds of responding "strongly agree" to teen birth control over "strongly disagree" are 13.1 times higher if the subject responded that premarital sex was "Not wrong at all" vs responding that it was "always wrong".

$$e^{\hat{\beta}(u_4-u_1)(v_4-v_1)} = e^{0.286\times(4-1)\times(4-1)} = 13.1$$

```
> oddrat1<-exp(coef(fit)["u:v"]*(4-1)*(4-1));oddrat1
     u:v
13.09878
```

Test of Independence

– Before, we used

$$X^2(I) = \sum \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}$$

to test independence.

– Conditional Test (LRT)

$$G^2(I|U) = G^2(I) - G^2(U), \text{ which is } \chi^2, \ df = 1.$$

– Wald Statistics

$$W = \left(\frac{\hat{\beta}}{SE_{\hat{\beta}}}\right)^2, \text{ is also } \chi^2, \ df = 1.$$

For ordinal data, these tests (2) and (3) are generally more powerful than (1).

Continue above statement, when X and Y are dependent, we are more likely to reject independence using $G^2(I|V)$ (LRT) or W (Wald) than when using $\chi^2(I)$.

Example

M_1: UA model
M_2: (X, Y) model independence model.

The fit of the UA model, M_1, is already performed. The fit of the independence model, M_2, is

```
> fit.ind<-glm(y~sex+birth,family=poisson);summary(fit.ind)

Call:
glm(formula = y ~ sex + birth, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-4.54742  -2.58174   0.06713   1.65192   5.25756

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.74742    0.09622  38.947  < 2e-16 ***
sex2        -0.97679    0.12166  -8.029 9.84e-16 ***
sex3        -0.34460    0.09881  -3.488 0.000487 ***
sex4         0.50920    0.08051   6.325 2.54e-10 ***
birth2       0.18859    0.10723   1.759 0.078611 .
birth3       0.71184    0.09683   7.352 1.96e-13 ***
birth4       0.45655    0.10136   4.504 6.66e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 431.08  on 15  degrees of freedom
Residual deviance: 127.65  on  9  degrees of freedom
AIC: 232.33

Number of Fisher Scoring iterations: 5
```

LRT:

$$G^2(M_2|M_1) = 127.65 - 11.53 = 116.12$$

with df=1, P-value $<$ 0.0001. This gives strong evidence of association. Alternatively, for the LRT test, one can use the "anova" command from R.

```
> anova(fit.ind,fit,test="Chi")
Analysis of Deviance Table

Model 1: y ~ sex + birth
Model 2: y ~ sex + birth + u:v
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         9    127.653
2         8     11.534  1   116.12 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Wald Test:

$$Z^2 = \frac{\hat{\beta}^2}{SE^2_{\hat{\beta}}} = \left(\frac{0.286}{0.0282}\right)^2 = 102.4$$

Ordinal Variables in Multi-way Table

Our method for two-way tables can be extended to higher order tables. If X, Y, Z are ordinal, we could fit the model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 ij + \beta_2 ik + \beta_3 jk. \quad (1)$$

This is a special case of the model (XY, XZ, YZ) where interaction terms λ_{ij}^{XY}, λ_{ik}^{XZ}, and λ_{jk}^{YZ} have been replaced by simpler terms exploiting ordinality.

If X and Y are ordinal but Z is nominal, then

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta ij + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \quad (2)$$

this is another special case of (XY, XZ, YZ). For both (1) and (2), we have uniform association in the XY partial table.

So conditional independence corresponds to $\beta_1 = 0$ in (1) or $\beta = 0$ in (2).

Alcohol, cigarette, marijuana example revisited

We can change our fit to match the output from SAS, which is shown in the textbook on Table 7.6, page 211. Basically, this fit is showing the estimates at the first levels of the variables instead of the second levels. By the manner in which the data was set up, we see that the coding is 0 = Yes and 1 = No. We want 1 = Yes and 0 = No. We don't have to redefine the table. It can be done easily in R as follow:

```
> fit.sas<-update(fit.Ha,contrasts=list(alcohol=as.matrix(c(1,0)),marijuana=as.m
+ cigarette=as.matrix(c(1,0))))
> summary(fit.sas,cor=F)

Call:
glm(formula = count ~ .^2, family = poisson, data = table7.3,
    contrasts = list(alcohol = as.matrix(c(1, 0)), marijuana = as.matrix(c(1,
        0)), cigarette = as.matrix(c(1, 0))))

Deviance Residuals:
       1         2         3         4         5         6         7         8
 0.02044  -0.02658  -0.09256   0.02890  -0.33428   0.09452   0.49134  -0.03690

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            5.63342    0.05970  94.361  < 2e-16 ***
marijuana1            -5.30904    0.47520 -11.172  < 2e-16 ***
cigarette1            -1.88667    0.16270 -11.596  < 2e-16 ***
alcohol1               0.48772    0.07577   6.437 1.22e-10 ***
marijuana1:cigarette1  2.84789    0.16384  17.382  < 2e-16 ***
marijuana1:alcohol1    2.98601    0.46468   6.426 1.31e-10 ***
cigarette1:alcohol1    2.05453    0.17406  11.803  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2851.46098  on 7  degrees of freedom
Residual deviance:    0.37399  on 1  degrees of freedom
AIC: 63.417

Number of Fisher Scoring iterations: 4
```

Fit of the saturated model on Relationship between College Enrollment and Political Affiliation

```
> mod.fitsat<-glm(formula = counts ~ enroll + pol_aff + enroll:pol_aff, data = d
+ family = poisson(link = log), na.action = na.exclude, control = list(epsilon =
Deviance = 4.440843e-08 Iterations - 1
Deviance = -6.667444e-15 Iterations - 2
> summary(mod.fitsat)

Call:
glm(formula = counts ~ enroll + pol_aff + enroll:pol_aff, family = poisson(link
    data = data2, na.action = na.exclude, control = list(epsilon = 1e-04,
        maxit = 50, trace = T))

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               3.52636    0.17150  20.562  < 2e-16 ***
enrolleng                -0.09237    0.24833  -0.372  0.70991
enrollagric              -0.58192    0.28643  -2.032  0.04219 *
enrolleduc               -0.39087    0.26998  -1.448  0.14768
pol_affdem                0.58451    0.21402   2.731  0.00631 **
pol_affindep             -0.75377    0.30317  -2.486  0.01291 *
enrolleng:pol_affdem     -1.07406    0.36152  -2.971  0.00297 **
enrollagric:pol_affdem   -0.39346    0.37671  -1.044  0.29628
enrolleduc:pol_affdem    -0.05645    0.33900  -0.167  0.86776
enrolleng:pol_affindep    0.15300    0.42777   0.358  0.72060
enrollagric:pol_affindep  0.58192    0.45502   1.279  0.20093
enrolleduc:pol_affindep   0.10318    0.46767   0.221  0.82538
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  6.9607e+01  on 11  degrees of freedom
Residual deviance: -6.6674e-15  on  0  degrees of freedom
AIC: 83.906

Number of Fisher Scoring iterations: 2
```