# What to do today (Mar 20)?

1. *Introduction and Preparation*
2. *Analysis with Binary Variables (Chp 1-2)*
3. *Analysis with Multicategory Variables (Chp 3)*
4. *Analysis with Count Variables (Chp 4)*

## 5. Model Selection and Evaluation (Chp 5)

- **5.1 Variable selection (Chp 5.1.1-4)**
- **5.2 Tools to assess model fit (Chp 5.2)**
- **5.3 Examples**

**To Return Marked Midterm 2 Papers**

6. *Additional Topics (Chp 6)*

# 5D. Model Selection and Evaluation: in loglinear regression

e.g. loglinear model for three-way contingency tables:

**Recall** that

- how to establish the association of the cell counts, $N_{ijk} \sim Poisson(\mu_{ijk})$, with $X$, $Y$, and $Z$, three categorical variables?

**Saturated Loglinear Model (XYZ)** (including all main effects, two factor interactions, three factor interactions: df=IJK)

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

**Loglinear Model of Multual Independence (X,Y,Z)** (including only main effects: df =I+J+K-2)

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

**Loglinear Model of Homogeneous Association (XY,YZ,XZ)** (including all main effects, two factor interactions: assuming $\lambda_{ijk}^{XYZ} = 0$)

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

Parameter Interpretation for Model (XY,YZ,XZ): when I=J=2, X-Y conditional odds ratio at $Z = k$ for any $k$ is

$$\log \theta_{XY(k)} = \log \left( \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} \right) = [\lambda_{11}^{XY} + \lambda_{22}^{XY}] - [\lambda_{12}^{XY} + \lambda_{21}^{XY}]$$

$\Longrightarrow$ *Homogeneous Conditional Association of X-Y*
Further, if $\lambda_{ij}^{XY} = 0$,

- $\Longrightarrow$ Model (YZ,XZ)
- $\log \theta_{XY(k)} = 0$, for all $k \Longrightarrow X \perp Y | Z$

**Statistical Inference** with (the loglinear (Poisson) regression with 3 categorical predictors):

- Be careful with coding $X, Y, Z$
- Choice of models: e.g. (X,Y,Z), (X,YZ), (YZ,XZ), (XY,YZ,XZ), (XYZ)
- Variouse inference procedures:
  - Estm model parameters; estm $\mu_{ijk}$; estm OR
  - Model checking/comparison: Pearson's $\chi^2$-test, LRT-test

**Example.** Alcohol, Cigarette and Marijuana Use

| Alcohol Use (A) | Cigarette Use (C) | Marijuana Use (M) | |
|---|---|---|---|
| | | Yes | No |
| Yes | Yes | 911 | 538 |
| | No | 44 | 456 |
| No | Yes | 3 | 43 |
| | No | 2 | 279 |

*Source: a survey conducted in 1992 by the Wright State Univ. School of Medicine and the United Health Services in Dayton.*

Using *read.table* to read in data and *as.data.frame* to form it into R's data format, or

- $counts < -c(911, 44, 3, 2, 538, 456, 43, 279)$
- $A < -gl(2, 2, 8)$; $C < -gl(2, 1, 8)$; $M < -gl(2, 4, 8)$;
  $\#\#1 = yes, 2 = no$
- $ACM.data < -cbind(A, C, M, counts)$

Run R to fit different models with the data: for example,

- For Model (ACM)

$$tmp.out < -glm(counts \sim A * C * M, family = poisson);$$

- For Model (AC,CM,AM)

$$tmp2.out < -glm(counts \sim A * C + C * M + A * M, family = poisson);$$

- For Model (CM,AM)

$$tmp3.out < -glm(counts \sim C * M + A * M, family = poisson);$$

- For Model (AC,M)

$$tmp4.out < -glm(counts \sim A * C + M, family = poisson);$$

- For Model (A,C,M)

$$tmp5.out < -glm(counts \sim A + C + M, family = poisson);$$

Step 1. Fitted Values for Loglinear Models:

- Plug in the estm for the parameters in the models to attain the fitted values, or
- Use "*tmp.out*$*fitted*", for example

The fit for (AC,AM,CM) is close to the observed data, the same as the fitted values for (ACM).

Fitted Values for Loglinear Models:

| | | | Loglinear Model | | | | |
|---|---|---|---|---|---|---|---|
| A | C | M | (A,C,M) | (AC,M) | (AM,CM) | (AC,AM,CM) | (ACM) |
| Yes | Yes | Yes | 540.0 | 611.2 | 909.24 | 910.4 | 911 |
| | | No | 740.2 | 837.8 | 438.84 | 538.6 | 538 |
| | No | Yes | 282.1 | 210.9 | 45.76 | 44.6 | 44 |
| | | No | 386.7 | 289.1 | 555.16 | 455.4 | 456 |
| No | Yes | Yes | 90.6 | 19.4 | 4.76 | 3.6 | 3 |
| | | No | 124.2 | 26.6 | 142.16 | 42.4 | 43 |
| | No | Yes | 47.3 | 118.5 | 0.24 | 1.4 | 2 |
| | | No | 64.9 | 162.5 | 179.84 | 279.6 | 279 |

Step 2. To obtain estimates for what needed based on the analyses

- using the analysis outputs: the estms for the model parameters and their estimated standard errors

- using the fitted counts when applicable

e.g. the OR of alcohol use (A yes vs not) between cigarette use or not (C yes vs not)

- conditional on marijuana use (M=yes or not)
- marginal (regardless of M)

Step 3. Chi-Squared Goodness-of-Fit Tests: Loglinear Residuals

- $G^2[(AC, AM, CM)] = 2 \sum n_{ijk} \log(\frac{n_{ijk}}{\hat{\mu}_{ijk}})$
- $X^2[(AM, CM)] = \sum \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$.
  - e.g. the Pearson's residuals: $e_{ijk} = \frac{n_{ijk} - \hat{\mu}_{ijk}}{\sqrt{\hat{\mu}_{ijk}}}$
  - residual plots: e.g. scatter plot of $e_{ijk}$ vs $A$

Step 4. Model Selection: (backward elimination)

| Start: AIC=65.04 $counts \sim ACM$ | | | |
|---|---|---|---|
| | Df | Deviance | AIC |
| - A:C:M | 1 | 0.37399 | 63.417 |
| $<$ *none* $>$ | | 0.00000 | 65.043 |
| **Step: AIC=63.42** $counts \sim A + C + M + AC + AM + CM$ | | | |
| | Df | Deviance | AIC |
| $<$ *none* $>$ | | 0.37 | 63.42 |
| - A:M | 1 | 92.02 | 153.06 |
| - A:C | 1 | 187.75 | 248.80 |
| - C:M | 1 | 497.37 | 558.41 |

# Part V.2.2D for Three-Way Contingency Tables

Step 5. Tests about Partial Associations:

- The test statistic for testing $\lambda^{AC} = 0$ in (AC,AM,CM) is

$$\begin{aligned} G^2[(AM, CM)\big|(AC, AM, CM)] &= G^2(AM, CM) - G^2(AC, AM, CM) \\ &= 187.8 - 0.04, \end{aligned}$$

df=2-1
$\Longrightarrow p < 0.001$: strong evidence against the null hypothesis and in favor of an A-C partial association.

# What to study next?

1. *Introduction and Preparation*
2. *Analysis with Binary Variables (Chp 1-2)*
3. *Analysis with Multicategory Variables (Chp 3)*
4. *Analysis with Count Variables (Chp 4)*
5. *Model Selection and Evaluation (Chp 5)*

6. **Additional Topics (Chp 6)**
   - **6.1 Exact inference (Chp 6.2)**
   - **6.2 Revisit to Loglinear and Logistic Models for Contingency Tables: the Loglinear-Logit Connection** (Supplementary)
   - *6.3 Revisit II to GLM and Some Advanced Topics (Chp 5.3, Chp 6.5)*