Topic 8: Log-Linear Models for Contigency Tables

Nov27

Questions to be solved

1. pp78 homogeneous association model.
2. Pp89 90

Checking Goodness of Fit Grouped vs Ungrouped data

1. **Explanatory Vars are solely categorical**
    1. residual df = number of paras in saturated model(number of setting s of x) - number of paras in the model
    2. Fixed number settings of predictor values is referred to as grouped data
2. **Explanatory vas are not solely categorical** (one can have continuous vars)
    1. Saturated Model: deviance does not necessarily follow a chi-square dist'n since the number of parameters is not fixed
    2. Ungrouped data, saturated model has a parameter for each subject
3. **Horseshoe crab data** if we group them
    1. Width has 66 unique values 66x2 contingency table
    2. Most fitted counts are very small
    3. when new data comes, additional with values would occur -> table dimension would grow(not fixed)

Log-Linear Models for 2 way tables

1. GLM using log link with poisson response-> model contigency table
2. $\pi_{ij} = \pi_{i+}\pi_{+j}$ hence $\mu_{ij} = n\pi_{i+}\pi_{+j}$
3. Loglinear Model of independence $log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$ with row effect λ_i^X and column effect λ_j^Y. Null Hypothesis of independence between two categorical variables is that model holds.
4. **I x 2 table**:
    1. $logit[P(Y=1|X=i)] = log\frac{P(Y=1|X=i)}{P(Y=2|X=i)} = log\frac{\mu_{i1}}{\mu_{i2}} = log\mu_{i1} - log\mu_{i2} = (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) = \lambda_1^Y - \lambda_2^Y$
    2. final term does not depend on i -> logit P Y = 1 X = i is identical at each level of X -> $logit[P(Y=1|X=i)] = \alpha$
    3. Odds of response in col1 equal $exp(\alpha) = exp(\lambda_1^Y - \lambda_2^Y)$ eg belief Yes estimated 1.49, odds of belief in the afterlife is exp(1.49) = 4.5 for each race
5. Saturated Loglinear Model: $log\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$
6. **Interpretation of Interaction**: $log\theta = log(\mu_{11}\mu_{22}\mu_{12}\mu_{21}) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$
7. Test of independence analyze whether these (I-1)(J-1) parameters equal to 0 residual df = (I-1)(J-1)
8. $\pi_{ij} = \frac{exp(lambda.ij..)}{\sum_a \sum_b lambda.a.b..}$

3-Way Tables

1. **Mutual Independence**:
    1. $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ for all ijk
    2. $log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$
2. **Joint Independence**: Y is **JI** of X and Z
    1. $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$ for all ijk
    2. $log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$
3. **Conditional Independence**: Conditional inde of X and Y, given Z
    1. $\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$ for all ijk
    2. $log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
4. μ_{ijk}: cell expected frequencies in the contigency table

    Single factor term in loglinear models for μ_{ijk} represent marginal distributions

    eg. inlclude lambda X in the model forces the fitted values to have the same totals at the various levels of X as do the observed data.

5. **Partial Association Models**:
    1. $log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$: **Homogeneous association model**: the conditional odds ratios between any two variables are identical at each level of the third variable (XY, YZ, XZ)
    2. $log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$: (XYZ), odds ratio between any two vars to vary across levels of the thrid variable

        Perfect Fit in a three-way table

6. Conditional Association:
    1. AM conditonal association for model AC AM CM

        $OR_{AM} = \frac{(\hat{\mu}_{A=Y,C=Y,M=Y})(\hat{\mu}_{A=N,C=Y,M=N})}{(\hat{\mu}_{A=Y,C=Y,M=N})(\hat{\mu}_{A=N,C=Y,M=Y})}$

        $OR_{AM} = \frac{(\hat{\mu}_{A=Y,C=N,M=Y})(\hat{\mu}_{A=N,C=N,M=N})}{(\hat{\mu}_{A=Y,C=N,M=N})(\hat{\mu}_{A=N,C=N,M=Y})}$

        $exp(\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}) = exp(\lambda_{11}^{XY})$

        use constraints for which parameters at the second level of any variable equal = 0

7. **Marginal Association**:
    1. AC marginal association for model AM CM

        $OR_{AM} = \frac{(\hat{\mu}_{A=Y,C=Y,M=Y+N})(\hat{\mu}_{A=N,C=N,M=Y+N})}{(\hat{\mu}_{A=Y,C=N,M=Y+N})(\hat{\mu}_{A=N,C=Y,M=Y+N})}$

    2. Compared with (AM,CM) and (ACM)model, the fit model

Model Checking and inference for log-linear models

1. **Fitting Log-Linear Models**
    1. $\hat{\mu}_{ijk} = \frac{n_{i+k}n_{+jk}}{n_{++k}}$ for model(XZ,YZ) of X-Y conditional independence
2. **Goodness of fit**
    1. $G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}log(\frac{n_{ijk}}{\hat{\mu}_{ijk}})$

   2. $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$
   3. Df = number of cell counts - number of non-redundant parameters
   4. saturated model has d.f. = 0 eg the (ACM)
3. **Residuals: adjusted or Pearson** tells why a particular model does not fit well or highlight cells that display lack of fit.

   Abs values of Adjusted Residuals:

   1. larger than 2 when there are few cells
   2. larger than 3 when there are many cells indicate lack of fit

4. **Partial Association**:

   (AC, AM, CM): null hypothesis of no partial association between alcohol use and cigarette smoking states that λ^{AC} term equals zero.

   that is: Test if the simpler model (AM, CM) o f A-C conditional independence holds against the alternative that (AC, AM, CM) holds.

5. **Likelihood Ratio Stat** $-2(L_0 - L_1)$ == G^2 stat, the df = diff between two df values

   eg. testing $\lambda^{AC} = 0$ in model (AC, AM, CM) is difference :

   $G^2(AM,CM) - G^2 (AC, AM, CM)) = 187.4 df = 2-1 = 1

   Small pvalue provides strong evidence agianst null hypothesis and in favor of an A-C partial association, so as other comparisons (AC, CM), (AC, AM) with AC AM Cm model. so we should use model (AC AM, CM) rather than any simpler models

6. **CI for Odds Ratios**:

   1. Use the estimate along with the standard errors to construct CI for true log odds ratios and then exponentiate them to tform intervals for odds ratios.
   2. Estimate conditional odds ratio between alcohol use and cigarette use
   3. $\hat{\lambda}_{11}^{AC}$ = 2.054 and ASE = 0.174
   4. 95 % CI for true conditional log odds ratios = 2.054 $\pm$ 1.96 $\times$ (0.174) = (1,71, 2.39)
   5. exp(1.71, 2.39) = (5.5, 11.0)
   6. A-M (8.0, 49.2) C-M (12.5, 23.8) intervals are wide but associations also strong.
   7. **There is a strong tendency for users of one drug to be useres of a second drug, and this is true both for users and nonusers of the third drug.**

Applied Corner to be updated.