What to do today (Mar 6)?

4. Analysis with Count Response (Chp 4)

4.1 Poisson Model for Counts (Chp 4.1)

4.2 Poisson Regression Analysis (Chp 4.2)

4.2.1 Introduction to Poisson regression models

4.2.2 Inference with Poisson regression models

4.2.3 Categorical explanatory variables

4.2.4 Poisson regression with contingency tables

4.3 Additional Topics on Count Responses (Chp 4.3-4)

4.3.1 Poisson rate regression

*4.3.2 Overdispersion and zero inflation**

4.3.3 Generalized linear models II

Revisit to **Example** of Belief in Afterlife:

	Belief in Afterlife	
Gender	Yes	No or Undecided
Females	435	147
Males	375	134

Analysis 1. Model of Independence $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$, corresponding to $\log \mu_{ij} = \lambda + \beta^X A + \beta^Y B$ with $\lambda_i^X = \beta^X A$ and $\lambda_j^Y = \beta^Y B$ and A, B the dummy variables.

Parameter	Coding Type 1	Coding Type 2	Coding Type 3
λ	4.876	6.069	5.472
λ_1^X	0.134	0	0.067
λ_2^X	0	- 0.134	-0.067
λ_1^Y	1.059	0	0.529
λ_2^Y	0	-1.059	-0.529

Analysis 2. Saturated Model $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$, corresponding to $\log \mu_{ij} = \lambda + \beta^X A + \beta^Y B + \beta^{XY} AB$ with $\lambda_i^X = \beta^X A$, $\lambda_j^Y = \beta^Y B$, and $\lambda_{ij}^{XY} = \beta^{XY} AB$, and A, B the dummy variables.
The number of nonredundant parameters:

$$1 + (I-1) + (J-1) + (I-1)(J-1) = IJ,$$

the same as the number of parameters as the $I \times J$ table has Poisson observations $\Longrightarrow$ perfect fit.

Association Parameter	Coding Type 1	Coding Type 2	Coding Type 3
λ_{11}^{XY}	0.056	0	0.014
λ_{12}^{XY}	0	0	-0.014
λ_{21}^{XY}	0	0	- 0.014
λ_{22}^{XY}	0	0.056	0.014

4.2.4B Poisson regression with contingency tables: Three-Way Contingency Tables

Recall that

- n individuals cross-classified according to X, Y, Z variables $\Longrightarrow$ an $I \times J \times K$ contingency table with cell counts $\{N_{ijk} : i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K\}$
- 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)

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

the table's df=IJK=num of non-redundant parameters

4.2.4B Poisson regression with contingency tables: Three-Way Contingency Tables

Loglinear Model of Independence (X,Y,Z) (including only main effects, i.e. one factor effects)

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

multual independence model; the table's df=IJK $>$ num of non-redundant parameters in the model 1+(I-1)+(J-1)+(K-1)

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

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

the table's df=IJK$>$num of non-redundant parameters

4.2.4B Poisson regression with contingency tables: Three-Way Contingency Tables

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

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

Thus, if $\lambda_{ij}^{XY} = 0$,

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

4.2.4B Poisson regression with contingency tables: Three-Way Contingency Tables

Statistical Inference

the statistical analysis with the loglinear (Poisson) regression model with three 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: the main effects, and/or two/three factor interactions
- Estm μ_{ijk}, and then OR
- Model checking/Comparison: Pearson's χ^2-test, LRT-test

e.g. H_0: Model (YZ,XZ) vs H_1: Model (XY,YZ,XZ)

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.

Step 1. Fitted Values for Loglinear Models: (software available to do so) 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.

- the A-C association with model (AM,CM):
  - Estimate of the conditional OR?
  - Estimate of the marginal OR?

- model (AC,AM,CM) permits all pairwise associations but maintains homogeneous odds ratios between two variables at each level of the third variable.
  - The A-C estimated conditional odds ratios for this model?
  - The A-C estimated marginal odds ratio?

Step 3. Confidence Intervals for Odds Ratios:
MLE of loglinear model parameters have large-sample normal distributions: to use the estimates and their ASE to construct confidence intervals for true log odds ratios and then exponentiate them to form intervals for odds ratios.
For example, in (AC,AM,CM)

- R: $\hat{\lambda}_{22}^{AC} = 2.054$ $(ASE = 0.174)$
- SAS - PROC GENMOD: $\hat{\lambda}_{11}^{AC} = 2.054$ $(ASE = 0.174)$
- SAS - PROC CATMOD: $\hat{\lambda}_{11}^{AC} = \hat{\lambda}_{22}^{AC} = 0.514$,

all $\Longrightarrow \hat{\lambda}_{11}^{AC} + \hat{\lambda}_{22}^{AC} - \hat{\lambda}_{12}^{AC} - \hat{\lambda}_{21}^{AC} = 2.054$ $(ASE = 0.174)$:
$\Longrightarrow$ 95% CI for log odds ratio: $2.054 \pm 1.96(0.174)$, yielding $(e^{1.71}, e^{2.39}) = (5.5, 11.0)$ for CI of the odds ratio.

Model Checking?

4.3.1A Poisson rate regression: Introduction

A "rate" variable is often of interest Y/t: e.g.

- number of computer crashes in some area
- number of arrivals at an airport over some time periods

When the baseline measure of the "exposure" varies over observations?

- The measure needs to be incorporated into the analysis.
- One way to do this is to model Y/t instead of just Y: Y =count of events; t =measure of opportunity for events.

Poisson Rate Regression Model. Consider the response $Y|t, x_1, \ldots, x_K \sim Poisson\big(\mu(x_1, \ldots, x_K; t)\big)$ and assume

$$\log[\mu(x_1, \ldots, x_K; t)] = \log(t) + \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K,$$

equivalently to $E(Y/t|\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K)$.

4.3.1A Poisson rate regression: Introduction

Poisson Rate Regression Model.
Consider the response $Y|t, x_1, \ldots, x_K \sim Poisson\big(\mu(x_1, \ldots, x_K; t)\big)$ and assume

$$\log[\mu(x_1, \ldots, x_K; t)] = \log(t) + \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K.$$

- $\log(t)$ is an offset: t helps to adjust the "usual" mean by the baseline measure.

$$E(Y|\mathbf{x}, t) = t \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K)$$

- *Statistical Inference.* estimation, testing, and model interpretation proceed in a similar manner as before.

4.3.1B Poisson rate regression: Example

Example. Number of Credit Cards vs Income
(*https://onlinecourses.science.psu.edu/stat504/node/170*)

Income[a]	Number Cases	Credit Cards
24	1	0
27	1	0
28	5	2
29	3	0
...	...	...
120	6	6
130	1	1

[a] in millions of lira
(the currency in Italy before euro)

Consider $\log(\mu/t) = \beta_0 + \beta_1$income:
$\mu = E$(number of credit cards), $t =$ number of cases.

4.3.1B Poisson rate regression: Example

- The fitted model:

$$\log(\hat{\mu}/t) = -2.3866 + 0.0208 \times \text{income}$$

  where $\log(t) = \log(\text{cases})$.

- Questions can be answered by the analysis:
  - What is the estimated average rate of incidence, i.e. the usage of credit cards given the income?
  - Is income a significant predictor? Does the overall model fit?

e.g. with income$= 65$,

$$\log(\hat{\mu}/t) = -2.3866+0.0208\times 65 \Longrightarrow \log(\hat{\mu}) = -2.3866+0.0208\times 65+\log(t)$$

for a group of six people

$$\log(\hat{\mu}) = -2.3866 + 0.0208 \times 65 + \log(6) \Longrightarrow \hat{\mu} = 2.126$$

What will we do next?

4. **Analsyis with Count Response (Chp 4)**
    - *4.1 Poisson Model for Count Data (Chp 4.1)*
    - *4.2 Poisson Regression Analysis (Chp 4.2)*
    - **4.3 Additional Topics on Count Responses (Chp 4.3-4)**
        - *4.3.1 Poisson rate regression*
        - **4.3.2 Zero inflation**
        - **4.3.3 Generalized linear models**
5. **Model Selection and Evaluation (Chp 5)**
6. *Additional Topics (Chp 6)*