Loglinear models

STAT 526
Professor Olga Vitek

April 19, 2011

Can Use Poisson Likelihood To Model Both Poisson and Multinomial Counts

Recall: Poisson Distribution

- Probability distribution:
  - Y - number of events in a fixed interval of space/time
  - $Y \sim Poisson(\lambda)$
  - $p(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \; y = 0, 1, \ldots; \quad E(Y) = var(Y) = \lambda$
  - $Y_1, Y_2, \ldots, Y_c \overset{ind}{\sim} Poisson(\lambda_i), \; \sum_{i=1}^{c} Y_i \sim Poisson(\sum_{i=1}^{c} \lambda_i)$
- c indep. Poisson r.v. $|$ total $\sim$ Multinomial

$$P(Y_1 = n_1, \ldots, Y_c = n_c \mid \sum_i Y_i = n)$$

$$= \frac{P(Y_1 = n_1, \ldots, Y_c = n_c)}{P(\sum_i Y_i = n)}$$

$$= \frac{\prod_i [\; exp(-\lambda_i)\lambda_i^{n_i}/n_i! \;]}{exp(-\sum_i \lambda_i) \; (\sum_i \lambda_i)^n/n!} = \frac{n!}{\prod_i n_i!} \prod_i \pi_i^{n_i}, \; \pi_i = \frac{\lambda_i}{\sum_i \lambda_i}$$

Models for Joint Distributions of Unordered Categorical Variables

Joint Distributions of Categorical Variables

- Convenient to model contingency tables
  - two-way, but also more complex tables
  - can express complex probabilistic relationships
- Treat all categorical variables symmetrically
  - no distinction between predictor and response
  - analogous to correlation of continuous variables
- Can be thought of as a model for a network of associations between categorical variables
  - related to graphical models
  - Example: modeling functional networks of genes (e.g. "activated/non-activated"; substitutions of nucleotides)

2 R.V.: Independence

- $Y_{ij} \sim Poisson(\lambda_{ij})$
  - Of interest: effect of row i and column j on Y_{ij}
- General model **(Faraway Ch. 4)**:

$$log\, E\{Y_{ij}\} \overset{notation}{=} log\, \lambda_{ij} = log\, n\, \pi_{ij}$$

- Assuming independence of rows and columns:

$$\begin{aligned} log\, E\{Y_{ij}\} &= log\, n\, \pi_i\, \pi_j = log\, n + log\, \pi_i + log\, \pi_j \\ \overset{notation}{=} & \quad \mu + \alpha_i + \beta_j, \\ where & \quad \sum_i e^{\alpha_i} = \sum_j e^{\beta_j} = 1 \end{aligned}$$

- ML estimation with Poisson likelihood

$$\begin{aligned} \hat{\pi}_{ij} &= e^{\hat{\alpha}_i} \cdot e^{\hat{\beta}_j} = \hat{\pi}_i\, \hat{\pi}_j; \\ \hat{\lambda}_{ij} &= n\, \hat{\pi}_{ij} \end{aligned}$$

  - Total number of parameters $1+(I-1)+(J-1)$
  - Same $\hat{\lambda}_{ij}$ as in X^2 test for independence
  - Can use X^2 and G^2 tests to test goodness of fit

2 R.V.: Independence

- Alternative parametrization **(Agresti Ch.8)**

$$log\,E\{Y_{ij}\} \;\; = \mu + \alpha_i + \beta_j,\;\; \alpha_I = \beta_J = 0$$

  - $\mu = log\,E\{Y_{IJ}\}$
  - α_i and β_j are deviations of $E\{Y_{ij}\}$ from the reference cell (I, J) due to row i and column j

- Parametrization in R
  - $\alpha_1 = \beta_1 = 0$ and $\mu = log\,E\{Y_{11}\}$
  - Will use this parametrization from now on.

- ML estimation with Poisson likelihood

$$\begin{aligned}
\widehat{\lambda}_{ij} &= e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j} \overset{equivalent}{=} n_{i+}n_{+j}/n \\
\widehat{\pi}_{ij} &\overset{Slide\;\;2}{=} \;=\frac{\lambda_{ij}}{\sum_i \sum_j \lambda_{ij}} = \frac{e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j}}{\sum_i \sum_j e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j}} \\
&= \frac{e^{\widehat{\alpha}_i}}{\sum_i e^{\widehat{\alpha}_i}} \cdot \frac{e^{\widehat{\beta}_j}}{\sum_j e^{\widehat{\beta}_j}} = \widehat{\pi}_i\,\widehat{\pi}_j \overset{equivalent}{=} n_{i+}n_{+j}/n^2
\end{aligned}$$

- As in ANOVA, all parametrizations produce identical estimates of probabilities and counts

2 R.V.: Saturated model

- Saturated model

$$
\begin{aligned}
log\, E\{Y_{ij}\} &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij} \\
\mu &= log\, E\{Y_{11}\},\ \alpha_1 = \beta_1 = \alpha\beta_{i1} = \alpha\beta_{1j} = 0
\end{aligned}
$$

  - Total number of parameters
    $1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$
    (i.e. describes each cell perfectly)

- ML estimation with Poisson likelihood

- Model diagnostics developed for Poisson regression (e.g. residuals) apply

- Test for independence of rows and columns

  - $H_0 : \alpha\beta_{ij} = 0$ vs $H_0 : \alpha\beta_{ij} \neq 0$

  - LR (G^2) test with $(I-1)(J-1)$ df

3 R.V.: Mutual Indep.

- 3-way $I \times J \times K$ cross-classification of r.v. X, Y and Z

- Assume the count $Y_{ijk} \sim Poisson(\, E\{Y_{ijk}\}\,)$

- X, Y and Z are mutually independent if $\pi_{ijk} = \pi_i \, \pi_j \, \pi_k$
  - $log\, E\{Y_{ijk}\} = log\, n + log\, \pi_i + log\, \pi_j + log\, \pi_k$
  - Total number of parameters $1 + (I-1) + (J-1) + (K-1)$

- The log-linear model is

$$\begin{aligned} log\, E\{Y_{ijk}\} &= \mu + \alpha_i + \beta_j + \gamma_k \\ \mu &= log\, E\{Y_{111}\},\ \alpha_1 = \beta_1 = \gamma_1 = 0 \end{aligned}$$

$$\begin{aligned} \hat{\lambda}_{ijk} &= e^{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\hat{\gamma}_k} \overset{equivalent}{=} n_{i++}n_{+j+}n_{++k}/n^2 \\ \hat{\pi}_{ijk} &\overset{Slide\ 2}{=} \frac{e^{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\hat{\gamma}_k}}{\sum_i \sum_j \sum_k e^{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\hat{\gamma}_k}} = \frac{e^{\hat{\alpha}_i}}{\sum_i e^{\hat{\alpha}_i}} \cdot \frac{e^{\hat{\beta}_j}}{\sum_j e^{\hat{\beta}_j}} \cdot \frac{e^{\hat{\gamma}_k}}{\sum_k e^{\hat{\gamma}_k}} \\ &= \hat{\pi}_i \, \hat{\pi}_j \, \hat{\pi}_k \overset{equivalent}{=} n_{i++}n_{+j+}n_{++k}/n^3 \end{aligned}$$

Example: Female Smoking

- A survey of women by age, and a follow-up study 20 years later.

```
library(faraway)
data(femsmoke)

> head(femsmoke)
   y smoker dead   age
1  2    yes  yes 18-24
2  1     no  yes 18-24
3  3    yes  yes 25-34
4  5     no  yes 25-34
5 14    yes  yes 35-44
6  7     no  yes 35-44

ct3 <- xtabs(y~smoker+dead+age, femsmoke)
> ct3
, , age = 18-24
      dead
smoker yes  no
   yes   2  53
   no    1  61
, , age = 25-34
      dead
smoker yes  no
   yes   3 121
   no    5 152
..................................................
```

Example: Female Smoking

- Pearson X^2 test of mutual independence

```
> summary(ct3)
Call: xtabs(formula=y~smoker+dead+age, data=femsmoke)
Number of cases in table: 1314
Number of factors: 3
Test for independence of all factors:
Chisq = 790.6, df = 19, p-value = 2.140e-155
```

- Log-linear model with mutual independence

```
> fit1 <- glm(y~smoker+dead+age, femsmoke, family="poisson")
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67778    0.10702  25.021  < 2e-16 ***
smokerno     0.22931    0.05554   4.129 3.64e-05 ***
deadno       0.94039    0.06139  15.319  < 2e-16 ***
age25-34     0.87618    0.11003   7.963 1.67e-15 ***
age35-44     0.67591    0.11356   5.952 2.65e-09 ***
age45-54     0.57536    0.11556   4.979 6.40e-07 ***
age55-64     0.70166    0.11307   6.206 5.45e-10 ***
age65-74     0.34377    0.12086   2.844  0.00445 **
age75+      -0.41837    0.14674  -2.851  0.00436 **

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1193.9  on 27  degrees of freedom
Residual deviance:  735.0  on 19  degrees of freedom
AIC: 887.2
```

3 R.V.: Joint Independence

- Assume the count $Y_{ijk} \sim Poisson(\, E\{Y_{ijk}\}\,)$

- X and Y are dependent, but together they are independent of Z
  $\pi_{ijk} = \pi_{ij}\,\pi_k$
  - $log\, E\{Y_{ijk}\} = log\, n + log\, \pi_{ij} + log\, \pi_k$
  - Total number of parameters
    $1 + (IJ - 1) + (K - 1)$

- The log-linear model is

$$
\begin{aligned}
log\, E\{Y_{ijk}\} &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k \\
\mu &= log\, E\{Y_{111}\}, \\
\alpha_1 &= \beta_1 = \gamma_1 = \alpha\beta_{1j} = \alpha\beta_{i1} = 0
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\lambda}_{ijk} &\overset{Slide\ 2}{=} e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\alpha\beta}_{ij}+\widehat{\gamma}_k} \overset{equivalent}{=} n_{ij+}n_{++k}/n \\
\widehat{\pi}_{ijk} &= \frac{e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\alpha\beta}_{ij}+\widehat{\gamma}_k}}{\sum_i \sum_j \sum_k e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\alpha\beta}_{ij}+\widehat{\gamma}_k}} = \frac{e^{\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\alpha\beta}_{ij}}}{\sum_i \sum_j e^{\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\alpha\beta}_{ij}}} \cdot \frac{e^{\widehat{\gamma}_k}}{\sum_k e^{\widehat{\gamma}_k}} \\
&= \widehat{\pi}_{ij}\,\widehat{\pi}_k \overset{equivalent}{=} n_{ij+}n_{++k}/n^2
\end{aligned}
$$

Example: Female Smoking

- Model joint independence of `smoker` and `dead` from `age` (i.e. `smoker` and `dead` are dependent, but jointly independent of `age`).
  - Only a minor improvement of model fit

```
> fit2 <- glm(y~smoker*dead+age, femsmoke, family="poisson")
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.51582    0.12239  20.555  < 2e-16 ***
smokerno         0.50361    0.10743   4.688 2.76e-06 ***
deadno           1.15910    0.09722  11.922  < 2e-16 ***
age25-34         0.87618    0.11003   7.963 1.67e-15 ***
age35-44         0.67591    0.11356   5.952 2.65e-09 ***
age45-54         0.57536    0.11556   4.979 6.40e-07 ***
age55-64         0.70166    0.11307   6.206 5.45e-10 ***
age65-74         0.34377    0.12086   2.844  0.00445 **
age75+          -0.41837    0.14674  -2.851  0.00436 **
smokerno:deadno -0.37858    0.12566  -3.013  0.00259 **

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1193.9  on 27  degrees of freedom
Residual deviance:  725.8  on 18  degrees of freedom
AIC: 880
```

3 R.V.: Conditional Indep.

- $P\{X = i\}$ and $P\{Y = j\}$ are independent, given $Z = k$
  $\pi_{ij|k} = \pi_{i|k}\,\pi_{j|k}$
  - weaker than mutual or joint independence

- The joint probability is then
  $\pi_{ijk} = \pi_{i|k}\,\pi_{j|k}\,\pi_k = \frac{\pi_{ik}}{\pi_k} \cdot \frac{\pi_{jk}}{\pi_k} \cdot \pi_k = \frac{\pi_{ik}\pi_{jk}}{\pi_k}$
  - $log\, E\{Y_{ijk}\} = log\, n + log\, \pi_{ik} + log\, \pi_{jk} - log\, \pi_k$

- The log-linear model is

$$
\begin{aligned}
log\, E\{Y_{ijk}\} &= \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} \\
\mu &= log\, E\{Y_{111}\}, \\
\alpha_1 &= \beta_1 = \gamma_1 = \alpha\gamma_{1k} = \alpha\gamma_{i1} = \beta\gamma_{1k} = \beta\gamma_{j1} = 0
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\pi}_{ijk} &= \frac{e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\gamma}_k+\widehat{\alpha\gamma}_{ik}+\widehat{\beta\gamma}_{jk}}}{\sum_i \sum_j \sum_k e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\gamma}_k+\widehat{\alpha\gamma}_{ik}+\widehat{\beta\gamma}_{jk}}} \\
&= \frac{e^{\widehat{\alpha}_i+\widehat{\gamma}_k+\widehat{\alpha\gamma}_{ik}}}{\sum_i \sum_k e^{\widehat{\alpha}_i+\widehat{\gamma}_k+\widehat{\alpha\gamma}_{ik}}} \cdot \frac{e^{\widehat{\beta}_j+\widehat{\gamma}_k+\widehat{\beta\gamma}_{jk}}}{\sum_j \sum_k e^{\widehat{\beta}_j+\widehat{\gamma}_k+\widehat{\beta\gamma}_{jk}}} \cdot \frac{\sum_k \widehat{\gamma}_k}{e^{\widehat{\gamma}_k}}
\end{aligned}
$$

Example: Female Smoking

- Conditional indep. of (smoke, dead | age)

```
> fit3<-glm(y~smoker*age+dead*age,femsmoke,family="poisson")
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.34377    0.58563   0.587 0.557199
smokerno             0.11980    0.18523   0.647 0.517785
age25-34             0.91760    0.68737   1.335 0.181895
age35-44             1.95402    0.62882   3.107 0.001887 **
age45-54             2.84979    0.60950   4.676 2.93e-06 ***
age55-64             3.44819    0.59868   5.760 8.43e-09 ***
age65-74             3.00134    0.61023   4.918 8.73e-07 ***
age75+               2.22118    0.64799   3.428 0.000609 ***
deadno               3.63759    0.58490   6.219 5.00e-10 ***
smokerno:age25-34    0.11616    0.22078   0.526 0.598789
smokerno:age35-44   -0.01536    0.22749  -0.068 0.946172
smokerno:age45-54   -0.63063    0.23414  -2.693 0.007074 **
smokerno:age55-64   -0.06894    0.22643  -0.304 0.760765
smokerno:age65-74    1.15649    0.26427   4.376 1.21e-05 ***
smokerno:age75+      1.47413    0.35617   4.139 3.49e-05 ***
age25-34:deadno     -0.10756    0.68613  -0.157 0.875435
age35-44:deadno     -1.33977    0.62810  -2.133 0.032920 *
age45-54:deadno     -2.17125    0.61128  -3.552 0.000382 ***
age55-64:deadno     -3.17171    0.59999  -5.286 1.25e-07 ***
age65-74:deadno     -4.94977    0.61512  -8.047 8.49e-16 ***
age75+:deadno      -26.30450 5776.51889  -0.005 0.996367

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1193.938  on 27  degrees of freedom
Residual deviance:    8.327  on  7  degrees of freedom
AIC: 184.52
```

3 R.V.: Uniform (Homogeneous) Association

- For each level of one variable, same association of the other two variables
  $\pi_{ijk} = \pi_{ij}\,\pi_{jk}\,\pi_{ik}$
  - $log\,E\{Y_{ijk}\} = log\,n + log\,\pi_{ij} + log\,\pi_{jk} + log\,\pi_{ik}$
- The log-linear model is

$$\begin{aligned} log\,E\{Y_{ijk}\} &= \mu + \alpha_i + \beta_j + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta_{ij} \\ \mu &= log\,E\{Y_{111}\}, \\ \alpha_1 &= \beta_1 = \gamma_k = \alpha\gamma_{1k} = \alpha\gamma_{i1} = \beta\gamma_{1k} = \beta\gamma_{j1} \\ &= \alpha\beta_{1j} + \alpha\beta_{i1} = 0 \end{aligned}$$

- Not a saturated model, since no 3-way interaction

$$\begin{aligned} \widehat{\pi}_{ijk} &= \frac{e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\gamma}_k+\widehat{\alpha\gamma}_{ik}+\widehat{\beta\gamma}_{jk}+\widehat{\alpha\beta}_{ij}}}{\sum_i\sum_j\sum_k e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\gamma}_k+\widehat{\alpha\gamma}_{ik}+\widehat{\beta\gamma}_{jk}+\widehat{\alpha\beta}_{ij}}} \\ &= \frac{e^{\widehat{\alpha}_i+\widehat{\gamma}_k+\widehat{\alpha\gamma}_{ik}}}{\sum_i\sum_k e^{\widehat{\alpha}_i+\widehat{\gamma}_k+\widehat{\alpha\gamma}_{ik}}} \cdot \frac{e^{\widehat{\beta}_j+\widehat{\gamma}_k+\widehat{\beta\gamma}_{jk}}}{\sum_j\sum_k e^{\widehat{\beta}_j+\widehat{\gamma}_k+\widehat{\beta\gamma}_{jk}}} \cdot \frac{e^{\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\alpha\beta}_{ij}}}{\sum_i\sum_j e^{\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\alpha\beta}_{ij}}} \\ &= \widehat{\pi}_{ij}\,\widehat{\pi}_{jk}\,\widehat{\pi}_{ik} \end{aligned}$$

Interpretation of Uniform Association

- Constant odds ratios between the levels of two variables, for each level of the third variable
  - e.g. for $i = 1, 2,\ j = 1, 2$ and a given level k:

    $$log\, OR = log \frac{\lambda_{11k}\, \lambda_{22k}}{\lambda_{12k}\, \lambda_{21k}} = \alpha\beta_{11} + \alpha\beta_{22} - \alpha\beta_{12} - \alpha\beta_{21}$$

  - independent of k

- No easy way to estimate $\widehat{\lambda}_{ijk}$ and $\widehat{\pi}_{ijk}$ based on cell counts

Example: Female Smoking

```
> fit4 <- glm(y~smoker+age+dead+smoker:age+
          smoker:dead+dead:age,femsmoke,family="poisson")
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.54284    0.58736   0.924 0.355384
smokerno            -0.29666    0.25324  -1.171 0.241401
age25-34             0.92902    0.68381   1.359 0.174273
age35-44             1.94048    0.62486   3.105 0.001900 **
age45-54             2.76845    0.60657   4.564 5.02e-06 ***
age55-64             3.37507    0.59550   5.668 1.45e-08 ***
age65-74             2.86586    0.60894   4.706 2.52e-06 ***
age75+               2.02211    0.64955   3.113 0.001851 **
deadno               3.43271    0.59014   5.817 6.00e-09 ***
smokerno:age25-34    0.11752    0.22091   0.532 0.594749
smokerno:age35-44    0.01268    0.22800   0.056 0.955654
smokerno:age45-54   -0.56538    0.23585  -2.397 0.016522 *
smokerno:age55-64    0.08512    0.23573   0.361 0.718030
smokerno:age65-74    1.49088    0.30039   4.963 6.93e-07 ***
smokerno:age75+      1.89060    0.39582   4.776 1.78e-06 ***
smokerno:deadno      0.42741    0.17703   2.414 0.015762 *
age25-34:deadno     -0.12006    0.68655  -0.175 0.861178
age35-44:deadno     -1.34112    0.62857  -2.134 0.032874 *
age45-54:deadno     -2.11336    0.61210  -3.453 0.000555 ***
age55-64:deadno     -3.18077    0.60057  -5.296 1.18e-07 ***
age65-74:deadno     -5.08798    0.61951  -8.213  < 2e-16 ***
age75+:deadno      -27.31727 8839.01146  -0.003 0.997534

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1193.9378  on 27  degrees of freedom
Residual deviance:    2.3809  on  6  degrees of freedom
AIC: 180.58
```

3 R.V.: ML Estimation

- Joint Poisson probability of cell counts Y_{ijk}

$$\prod_i \prod_j \prod_k \frac{e^{-\lambda_{ijk}} \cdot \lambda_{ijk}^{n_{ijk}}}{n_{ijk}!}$$

- The log-likelihood

$$l(\mu) = \sum_i \sum_j \sum_k n_{ijk} log\, \lambda_{ijk} - \sum_i \sum_j \sum_k \lambda_{ijk} + C$$

- For the model with joint independence

$$\begin{aligned} l(\lambda) &= n\mu + \sum_i n_{i++}\alpha_i + \sum_j n_{+j+}\beta_j + \sum_k n_{++k}\gamma_k \\ &\quad - \sum_i \sum_j \sum_k e^{\mu+\alpha_i+\beta_j+\gamma_k} + C \end{aligned}$$

  – n_{i++}, n_{+j+}, n_{++k} are sufficient statistics
    parameters are estimated in these terms

Loglinear Models Summary

- Y_{ijk} - count in cell (i, j, k); $Y_{ijk} \sim Poisson(\lambda_{ijk})$
- Conditional on $n = \sum_{ijk} n_{ijk}$,

  $Y_{ijk} \sim Multinom(\pi_{ijk})$.
- $\mu_{ijk} = log\, E\{Y_{111}\}$ - reference cell
- α_i, β_j, γ_k - deviations of $log\, E\{Y_{ijk}\}$ from reference; $\alpha_1 = \beta_1 = \gamma_1 = 0$.
- Residual Df = $IJK - \#$ model params

Model	log $E\{Y_{ijk}\} =$
Mut. Indep	$\mu + \alpha_i + \beta_j + \gamma_k$
Joint Indep.	$\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$
Cond. Indep.	$\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$
Unif. Assoc.	$\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$
Saturated	$\mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$

Model	$\pi_{ijk} =$	$\hat{\lambda}_{ijk} =$
Mut. Indep	$\pi_i \pi_j \pi_k$	$n_{i++} n_{+j+} n_{++k} / n^2$
Joint Indep.	$\pi_{ij} \pi_k$	$n_{ij+} n_{++k} / n$
Cond. Indep.	$\pi_{ik} \pi_{jk} / \pi_k$	$n_{i+k} n_{+jk} / n_{++k}$
Unif. Assoc.	$\pi_{ij} \pi_{ik} \pi_{jk}$	Iterative
Saturated	π_{ijk}	n_{ijk}

Models for Joint Distributions of Ordered Categorical Variables

2 R.V.: Linear-by-Linear Association

- $Y_{ij} \sim Poisson(\lambda_{ij})$
  - Of interest: effect of row i and column j on Y_{ij}
  - Assign scores u_i to rows, $u_1 \leq u_2 \leq \ldots \leq u_I$
  - Assign scores v_j to columns, $v_1 \leq v_2 \leq \ldots \leq v_J$

- Log-linear model:

$$\begin{aligned} log\, E\{Y_{ij}\} &\overset{notation}{=} log\, \lambda_{ij} = log\, n\, \pi_{ij} \\ &= log\, n + \alpha_i + \beta_j + \gamma u_i v_j \end{aligned}$$

- The log-linear model:

$$\begin{aligned} log\, E\{Y_{ij}\} \quad & = \mu + \alpha_i + \beta_j + \gamma u_i v_j, \\ where \quad & \alpha_1 = \beta_1 = 0 \\ & \mu = log\, E\{Y_{11}\} - \gamma u_1 v_1, \\ & (= log\, E\{Y_{11}\} \text{ when } u_1 = v_1 \overset{coded\ as}{=} 0) \end{aligned}$$

  - γ - quantifies (positive or negative) association
  - Check sensitivity of conclusions to score coding

Interpretation of Linear-by-Linear Association

- Constant log-odds ratios for equally spaced scores
  - e.g. for adjacent entries in both rows and columns of the table:

    $$log\,OR = log\,\frac{\lambda_{ij}\,\lambda_{i+1,\,j+1}}{\lambda_{i,\,j+1}\,\lambda_{i+1,\,j}} = \gamma(u_{i+1} - u_i)(v_{j+1} - v_j)$$

  - same for non-adjacent equally-spaced scores

Example: Vote by Education

- Voting preference in 1996 pres. election

```
> library(faraway)
> data(nes96)
> xtabs(~PID+educ, nes96)
         educ
PID        MS HSdrop HS Coll CCdeg BAdeg MAdeg
  strDem    5     19 59   38    17    40    22
  weakDem   4     10 49   36    17    41    23
  indDem    1      4 28   15    13    27    20
  indind    0      3 12    9     3     6     4
  indRep    2      7 23   16     8    22    16
  weakRep   0      5 35   40    15    38    17
  strRep    1      4 42   33    17    53    25
```

- Introduce scores u_i and v_i

```
partyed$oPID <- unclass(partyed$PID)
partyed$oeduc <- unclass(partyed$educ)
> partyed
       PID   educ Freq oPID oeduc
1   strDem     MS    5    1     1
2  weakDem     MS    4    2     1
3   indDem     MS    1    3     1
4   indind     MS    0    4     1
5   indRep     MS    2    5     1
6  weakRep     MS    0    6     1
....................................
```

Example: Vote by Education

- Fit additive model (i.e. independence)

- Ignore potential order in the categories

```
> fit5 <- glm(Freq ~ PID + educ, partyed, family=poisson)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0131     0.2844   3.563 0.000367 ***
PIDweakDem   -0.1054     0.1027  -1.026 0.305125
PIDindDem    -0.6162     0.1194  -5.160 2.47e-07 ***
PIDindind    -1.6874     0.1790  -9.429  < 2e-16 ***
PIDindRep    -0.7550     0.1251  -6.038 1.56e-09 ***
PIDweakRep   -0.2877     0.1080  -2.663 0.007735 **
PIDstrRep    -0.1335     0.1035  -1.290 0.197038
educHSdrop    1.3863     0.3101   4.471 7.80e-06 ***
educHS        2.9485     0.2845  10.363  < 2e-16 ***
educColl      2.6662     0.2868   9.295  < 2e-16 ***
educCCdeg     1.9349     0.2967   6.521 6.98e-11 ***
educBAdeg     2.8600     0.2852  10.029  < 2e-16 ***
educMAdeg     2.2792     0.2912   7.827 4.99e-15 ***

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 626.798  on 48  degrees of freedom
Residual deviance:  40.743  on 36  degrees of freedom
AIC: 276.44
```

Example: Vote by Education

- Incorporate the quantitative scores

```
> fit6 <- glm(Freq~PID+educ+I(oPID*oeduc),
          partyed,family=poisson)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.245874   0.290541   4.288 1.80e-05 ***
PIDweakDem      -0.231690   0.109657  -2.113 0.034613 *
PIDindDem       -0.870935   0.142609  -6.107 1.01e-09 ***
PIDindind       -2.072665   0.214863  -9.646  < 2e-16 ***
PIDindRep       -1.272907   0.203997  -6.240 4.38e-10 ***
PIDweakRep      -0.940284   0.231496  -4.062 4.87e-05 ***
PIDstrRep       -0.922944   0.270247  -3.415 0.000637 ***
educHSdrop       1.288644   0.311257   4.140 3.47e-05 ***
educHS           2.749103   0.290065   9.478  < 2e-16 ***
educColl         2.360892   0.300152   7.866 3.67e-15 ***
educCCdeg        1.519473   0.321228   4.730 2.24e-06 ***
educBAdeg        2.330228   0.327217   7.121 1.07e-12 ***
educMAdeg        1.630806   0.353532   4.613 3.97e-06 ***
I(oPID * oeduc)  0.028745   0.009062   3.172 0.001513 **

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 626.798  on 48  degrees of freedom
Residual deviance:  30.568  on 35  degrees of freedom
AIC: 268.26
```

2 R.V.: Column-Effect Model

- $Y_{ij} \sim Poisson(\lambda_{ij})$
  - 'Ordinal-by-nominal' model (i.e. columns are not assigned scores)
  - Treat education as nominal, but political preference as ordinal with scores u_i, $u_1 \leq \ldots \leq u_I$

- Log-linear model:

$$\begin{aligned} log\, E\{Y_{ij}\} &\stackrel{notation}{=} log\, \lambda_{ij} = log\, n\, \pi_{ij} \\ &= log\, n + \alpha_i + \beta_j + \gamma_j u_i \end{aligned}$$

- The log-linear model:

$$\begin{aligned} log\, E\{Y_{ij}\} \quad &= \mu + \alpha_i + \beta_j + \gamma_j u_i, \\ where \quad &\alpha_1 = \beta_1 = \gamma_J = 0 \\ &\mu = log\, E\{Y_{11}\} - \gamma_1 u_1, \\ &(= log\, E\{Y_{11}\} \text{ when } u_1 \stackrel{coded\ as}{=} 0) \end{aligned}$$

  - γ_j - separate parameter of u_i for each column
  - $\widehat{\gamma}_j$ roughly equally spaced monotone if linear-by-linear model is appropriate

Constraints on γ_j in Column-Effect Model

- An example of a 2×3 table:

$$\begin{bmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \end{bmatrix}$$

- In matrix form, setting $u_1 = 1$ and $u_2 = 2$:

$$log\, E \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{bmatrix} = \left[\begin{array}{c|ccc|cc} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 0 \\ 1 & 1 & 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 & 0 & 0 \end{array}\right] \begin{bmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ \beta_3 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}$$

- α_1 and β_1 are constrained to avoid multicollinearity with intercept
- γ_3 is constrained to avoid multicollinearity with columns corresponding to μ and $\alpha 2$
  - constraint on γ_J ensures the same interpretation of the intercept?

Example: Vote by Education

- Column-Effect Model

```
> fit7<-glm(Freq~PID+educ+educ:oPID,partyed,family=poisson)
Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.945679   0.471322   4.128 3.66e-05 ***
PIDweakDem       -0.075007   0.109326  -0.686 0.492661
PIDindDem        -0.560399   0.140521  -3.988 6.66e-05 ***
PIDindind        -1.610437   0.210250  -7.660 1.86e-14 ***
PIDindRep        -0.660584   0.192522  -3.431 0.000601 ***
PIDweakRep       -0.178992   0.211862  -0.845 0.398193
PIDstrRep        -0.013421   0.241404  -0.056 0.955663
educHSdrop        1.061194   0.527977   2.010 0.044439 *
educHS            2.161235   0.484489   4.461 8.16e-06 ***
educColl          1.650803   0.491805   3.357 0.000789 ***
educCCdeg         0.971275   0.513874   1.890 0.058744 .
educBAdeg         1.722897   0.489151   3.522 0.000428 ***
educMAdeg         1.281529   0.501813   2.554 0.010655 *
educMS:oPID      -0.312217   0.154051  -2.027 0.042692 *
educHSdrop:oPID -0.194451   0.077228  -2.518 0.011806 *
educHS:oPID      -0.055347   0.048196  -1.148 0.250810
educColl:oPID     0.004460   0.050603   0.088 0.929760
educCCdeg:oPID   -0.008699   0.060667  -0.143 0.885978
educBAdeg:oPID    0.034554   0.048782   0.708 0.478740
educMAdeg:oPID          NA         NA      NA       NA
    Null deviance: 626.798  on 48  degrees of freedom
Residual deviance:  22.761  on 30  degrees of freedom
AIC: 270.46
```

Ordinal Models Summary

- Y_{ij} - count in row i, column j; $Y_{ij} \sim Poisson(\lambda_{ij})$
- Conditional on $\sum_i \sum_j n_{ij}$, $Y_{ij} \sim Multinomial(\pi_{ij})$.

Nominal Categories

- α_i, β_j - row and column effects, $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$

Ordinal Categories

- u_i, v_j - continuous scores of rows/columns γ or γ_i, γ_j - parameters, $\sum_{i=1}^{I} \gamma_i = \sum_{j=1}^{J} \gamma_j = 0$.

Model	log $E\{Y_{ij}\} =$	Residual Df
Independence	$\mu + \alpha_i + \beta_j$	$(I-1)(J-1)$
Linear-by-linear	$\mu + \alpha_i + \beta_j + \gamma(u_i v_j)$	$(I-1)(J-1)-1$
Row-effect	$\mu + \alpha_i + \beta_j + \gamma_i v_j$	$(I-1)(J-2)$
Column-effect	$\mu + \alpha_i + \beta_j + \gamma_j u_i$	(I-2)(J-1)
Saturated	$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	0

Multinomial Response as Function of Predictors: Surrogate Log Linear Models

Example: Housing

- Satisfaction of householders with housing **(Venable and Ripley Sec. 7.3)**

```
> head(housing)
     Sat   Infl  Type Cont Freq
1    Low    Low Tower  Low   21
2 Medium    Low Tower  Low   21
3   High    Low Tower  Low   28
4    Low Medium Tower  Low   34
5 Medium Medium Tower  Low   22
6   High Medium Tower  Low   36
.........................................


> xtabs(Freq~Infl+Type+Cont, data=housing)

, , Cont = Low
        Type
Infl     Tower Apartment Atrium Terrace
  Low       70       101     32      31
  Medium    92       118     28      41
  High      57        98     22      23
, , Cont = High
        Type
Infl     Tower Apartment Atrium Terrace
  Low       70       167     63      93
  Medium    80       179     56      65
  High      31       102     38      24
```

Null Model

- Model `Sat` as multinomial response; `Influence`, `Contact` and `Type` as predictors
  - A different offset per covariate pattern
  - Offset = 3-way interaction of all predictors
  - Same probabilities or response categories across all covariate patterns

$$
\begin{aligned}
log\, E\{Y_{ijkl}\} &= \mu_{ijk} + \delta_l \\
&= [\ \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \\
&\quad (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}\big]\ + \delta_l
\end{aligned}
$$

  - Constraints on $\alpha, \beta, \gamma, \delta$ and on the interactions

```
> fit <- glm(Freq ~ Infl*Type*Cont + Sat,
                              family=poisson, data=housing)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.136e+00  1.196e-01  26.225  < 2e-16 ***
InflMedium      2.733e-01  1.586e-01   1.723 0.084868 .
InflHigh       -2.054e-01  1.784e-01  -1.152 0.249511
TypeApartment   3.666e-01  1.555e-01   2.357 0.018403 *
TypeAtrium     -7.828e-01  2.134e-01  -3.668 0.000244 ***
TypeTerrace    -8.145e-01  2.157e-01  -3.775 0.000160 ***
.............................................................
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 833.66  on 71  degrees of freedom
Residual deviance: 217.46  on 46  degrees of freedom
```

Additive Contribution of Individual Predictors

- Testing whether `Sat` depends on each of the 3 predictors individually
  - A different offset per covariate pattern
  - The probabilities of response categories are affected by one predictor across all covariate patterns

$$
\begin{aligned}
log\, E\{Y_{ijkl}\} &= \mu_{ijk} + \delta_l + (\alpha\delta)_{il},\ or \\
log\, E\{Y_{ijkl}\} &= \mu_{ijk} + \delta_l + (\beta\delta)_{jl},\ or \\
log\, E\{Y_{ijkl}\} &= \mu_{ijk} + \delta_l + (\gamma\delta)_{kl}
\end{aligned}
$$

  - Constraints on $\alpha, \beta, \gamma, \delta$ and on the interactions

```
> addterm(fit, ~. + Sat:(Infl+Type+Cont),
    test="Chisq")
Single term additions
Model:
Freq ~ Infl * Type * Cont + Sat
         Df Deviance    AIC     LRT   Pr(Chi)
<none>         217.46 610.43
Infl:Sat  4    111.08 512.05 106.371 < 2.2e-16 ***
Type:Sat  6    156.79 561.76  60.669 3.292e-11 ***
Cont:Sat  2    212.33 609.30   5.126   0.07708 .
```

  - `Infl`: max reduction in resid. deviance & AIC

Additive Contributions of All Predictors

- Add main effects of all predictors
  - A different offset per covariate pattern
  - Same effect of each predictor on probabilities of response categories, regardless of the value of the other predictors
  - Constraints on $\alpha, \beta, \gamma, \delta$ and on the interactions

$$log\, E\{Y_{ijkl}\} = \mu_{ijk} + \delta_l + (\alpha\delta)_{il} + (\beta\delta)_{jl} + (\gamma\delta)_{kl}$$

```
fit1 <- update(fit, .~. + Sat:(Infl+Type+Cont))
Residual deviance:  38.662  on 34  degrees of freedom
```

- Add higher-order interactions to represent non-additive effects of predictors on `Sat`

$$\begin{aligned} log\, E\{Y_{ijkl}\} &= \mu_{ijk} + \delta_l + (\alpha\delta)_{il} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} \\ &\quad (\alpha\beta\delta)_{ijl} + (\alpha\gamma\delta)_{ikl} + (\beta\gamma\delta)_{jkl} \end{aligned}$$

```
addterm(fit1, .~.+Sat:(Infl+Type+Cont)^2, test="Chisq")
```

  - None significant

- Next analysis steps: plot predicted probabilities and counts; analysis of residuals

Compare to Multinom. Reg.

- Multinomial regression:

```
> library(nnet)
> fit.multinom <- multinom(Sat ~ Infl + Type + Cont,
         weights=Freq, data=housing)
Coefficients:
       (Intercept) InflMedium  InflHigh TypeApartment
Medium  -0.4192316  0.4464003 0.6649367    -0.4356851
High    -0.1387453  0.7348626 1.6126294    -0.7356261
       ........................
Residual Deviance: 3470.084
```

  - Different deviances due to different saturated models. In `multinom` the saturated model models subjects; in surrogate linear model it models covariate pattern
  - Can compare the predicted probabilities

    `p1 <- predict(fit.multinom, type="probs")`

- Saturated model with `multinom`

```
fit.saturated <- multinom(Sat ~ Infl * Type * Cont,
          weights=Freq, data=housing)
anova(fit.multinom, fit.saturated)
```

  - LR test statistic = residual deviance in surrogate linear model.

Models with Poisson Likelihood: Summary

- Model $E\{response\}$ as function of predictors
  - Count response:
    Poisson regression
  - Count response with overdispersion:
    Quasipoisson or Negative Binomial regression
  - Multinomial response:
    Surrogate log linear models

- Multivariate associations of categorical variables in contingency tables
  - Nominal random variables:
    Loglinear models
  - Ordinal random variables:
    Linear-by-linear model
    Row-effect and column-effect models