last change: 04.01.13@13.00

4.3 Multinomial Regression

Summary: multinomial logit, ML estimation, grouped data, log linear model equivalence, multinomial probit, multivariate probit

4.3.1 Multinomial Logit

Intro: J categrories, one indep variable, continuous case:

$$\mathbb{P}(Y = j \mid x) = \begin{cases} \frac{\exp(\alpha_j + \beta_j x)}{1+\sum_{j=2}^{J} \exp(\alpha_j+\beta_j x)} & \text{wenn } 1 < j \leq J \\ \frac{1}{1+\sum_{j=2}^{J} \exp(\alpha_j+\beta_j x)} & \text{wenn } j = 1 \end{cases}$$

with

Data

Y, X

	1	$\cdots$	J
x_1	n_{11}		n_{1J}
		$\ddots$	
x_r	n_{r1}		n_{rJ}

Model

Y, X

	1	$\cdots$	J
x_1	$p_{1\|x_1}$		$p_{J\|x_1}$
		$\ddots$	
x_r	$p_{1\|x_r}$		$p_{J\|x_r}$

parameters

Y	categories	1	2	$\cdots$	J
	parameters	$0,\ 0$	$\alpha_2,\ \beta_2$		$\alpha_J,\ \beta_J$

Motivation: Poisson case revisited
$\mu(x) = e^{\alpha+\beta x}$

$\mu(x)$

Y, X

	1	$\cdots$	J
x_1	$\mu_1(x_1)$		$\mu_J(x_1)$
		$\ddots$	
x_r	$\mu_1(x_r)$		$\mu_J(x_r)$

$p(x, y)$

Y, X

	1	$\cdots$	J
x_1	$p(x_1, 1)$		$p(x_1, J)$
		$\ddots$	
x_r	$p(x_r, 1)$		$p(x_r, J)$

$$\left.\begin{array}{l} \mu_j(x) = e^{\alpha_j+\beta_j x} \\ \mu_j(x) = \mu_0 p(x, j) \end{array}\right\} \Longrightarrow p(j|\, x) = \frac{p_{xj}}{p_{x\cdot}} = \frac{e^{\alpha_j+\beta_j x}}{\sum_j e^{\alpha_j+\beta_j x}}$$

(for identifiability choose $(Y = 1)$ as "baseline category" $(\alpha_1 = \beta_1 = 0)$)
hence

$$\frac{e^{\alpha_j+\beta_j x}}{\sum_j e^{\alpha_j+\beta_j x}} = \begin{cases} \frac{\exp(\alpha_j + \beta_j x)}{1+\sum_{j=2}^{J} \exp(\alpha_j+\beta_j x)} & \text{wenn } 1 < j \leq J \\ \frac{1}{1+\sum_{j=2}^{J} \exp(\alpha_j+\beta_j x)} & \text{wenn } j = 1 \end{cases}$$

Note:

1. equivalently in log-odds ratio as linear model

$$\log \frac{\mathbb{P}(Y = j \mid x)}{\mathbb{P}(Y = 1 \mid x)} = \alpha_j + \beta_j x$$

   or

$$\log \begin{bmatrix} p_{2|x} \big/ p_{1|x} \\ \vdots \\ p_{J|x} \big/ p_{1|x} \end{bmatrix} = \begin{bmatrix} \alpha_2 \\ \vdots \\ \alpha_J \end{bmatrix} + \begin{bmatrix} \beta_2 \\ \\ \beta_J \end{bmatrix} x$$

   and

$$p_{1|x} = \frac{1}{1 + p_{2|x} \; / \, p_{1|x} + \cdots + p_{J|x} \; / \, p_{1|x}}$$

2. interpretation $\beta_j > 0$ odds of j^{th} to 1^{st} category increasing with x
3. related log linear model ("Poisson Regression model") $\log \mu_j(x) = \alpha_j + \beta_j x$ independent regression models for $j = 1, \ldots, J$. In applications the logit model is much more important than the Poisson regression since the problem of occurrenc of one in J categories in dependence of observed covariate is of great practical importance.
4. for ML estimation the data may be individual $(x_1, y_1), \ldots, (x_n, y_n)$ with $y_1, \ldots, y_n \in \{1, \ldots, J\}$

Generalize: Several regressors: $X_1, \ldots, X_I$.

Y	categories	1	2	$\cdots$	J
	parameters	$0,\ 0, \ldots, 0$	$\alpha_2,\ \beta_{12}, \ldots, \beta_{I2}$		$\alpha_J,\ \beta_{1J}, \ldots, \beta_{IJ}$
x	covariates		$x_1, \ldots, x_I$		$x_1, \ldots, x_I$

Model

Motivation: general loglin model with more than one covariate
$\Longrightarrow$

$$\mathbb{P}(Y = j \mid \mathbf{x}) = \frac{\exp\left(\alpha_j + \sum_{i=1}^{I} \beta_{ij} x_i\right)}{\sum_{j=1}^{J} \exp\left(\alpha_j + \sum_{i=1}^{I} \beta_{ij} x_i\right)}$$

$$\alpha_1 = \beta_{i1} = 0 \text{ for } i = 1, \ldots, I$$

or

$$\mathbb{P}(Y = j \mid \mathbf{x}) = \begin{cases} \frac{\exp\left(\alpha_j + \sum_{i=1}^{I} \beta_{ij} x_i\right)}{1 + \sum_{j=2}^{J} \exp\left(\alpha_j + \sum_{i=1}^{I} \beta_{ij} x_i\right)} & \text{wenn } \; 1 < j \leq J \\ \frac{1}{1 + \sum_{j=2}^{J} \exp\left(\alpha_j + \sum_{i=1}^{I} \beta_{ij} x_i\right)} & \text{wenn } \; j = 1 \end{cases}$$

or, in short form

$$\log \frac{\mathbb{P}(Y = j \mid \mathbf{x})}{\mathbb{P}(Y = 1 \mid \mathbf{x})} = \alpha_j + \sum_{i=1}^{I} \beta_{ij} x_i$$

or

$$\log \begin{bmatrix} p_{2|\mathbf{x}} / p_{1|\mathbf{x}} \\ \vdots \\ p_{J|\mathbf{x}} / p_{1|\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \alpha_2 \\ \vdots \\ \alpha_J \end{bmatrix} + \sum_{i=1}^{I} \begin{bmatrix} \beta_{i2} \\ \vdots \\ \beta_{i2} \end{bmatrix} x_i$$

or

$$\log \begin{bmatrix} p_{2|\mathbf{x}} / p_{1|\mathbf{x}} \\ \vdots \\ p_{J|\mathbf{x}} / p_{1|\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \alpha_2 \\ \vdots \\ \alpha_J \end{bmatrix} + \begin{bmatrix} x_1 & \cdots & x_I \end{bmatrix} \begin{bmatrix} \beta_{12} & \cdots & \beta_{1J} \\ \vdots & \ddots & \vdots \\ \beta_{I2} & \cdots & \beta_{IJ} \end{bmatrix}$$

ParameterInterpretation: odds & odds ratio $(I = 1)$
interpret as **odds to reference group**:

$$\frac{\mathbb{P}\left(Y = j \mid x\right)}{\mathbb{P}\left(Y = 1 \mid x\right)} = e^{\alpha_j + \beta_j x}$$

interpret as **odds ratio**:

$$\frac{\mathbb{P}(Y = j \mid x + \Delta x)}{\mathbb{P}(Y = 1 \mid x + \Delta x)} \Big/ \frac{\mathbb{P}(Y = j \mid x)}{\mathbb{P}(Y = 1 \mid x)} = e^{\beta_j \Delta x}$$

interpret as **odds**:

general

$$\frac{\mathbb{P}(Y = j \mid x)}{\mathbb{P}(Y = i \mid x)} = e^{\alpha_j - \alpha_i + (\beta_j - \beta_i)x}$$

base line

$$\frac{\mathbb{P}(Y = j \mid x)}{\mathbb{P}(Y = 1 \mid x)} = \exp\left(\alpha_j + \beta_j x\right)$$

interpret as **odds ratio**:

$$\frac{\mathbb{P}(Y = j \mid x + \Delta x)}{\mathbb{P}(Y = i \mid x + \Delta x)} \Big/ \frac{\mathbb{P}(Y = j \mid x)}{\mathbb{P}(Y = i \mid x)} = e^{(\beta_j - \beta_i)\Delta x}$$

in words: "the odds of j to i are $e^{(\beta_j - \beta_i)\Delta x}$ times the odds of j to i if x increases by one unit Δx"

general

$$\frac{\mathbb{P}(Y = j \mid x + \Delta x)}{\mathbb{P}(Y = 1 \mid x + \Delta x)} \Big/ \frac{\mathbb{P}(Y = j \mid x)}{\mathbb{P}(Y = 1 \mid x)} = \exp\left(\beta_j \Delta x\right)$$

interpret **%wise change in odds ratio to reference group**: since $\exp\left(\beta_j \Delta x\right) \approx 1 + \beta_j' \Delta x$

$\Longrightarrow$

$$\frac{\frac{\mathbb{P}(Y=j \mid x+\Delta x)}{\mathbb{P}(Y=1 \mid x+\Delta x)} - \frac{\mathbb{P}(Y=j \mid x)}{\mathbb{P}(Y=1 \mid x)}}{\frac{\mathbb{P}(Y=j \mid x)}{\mathbb{P}(Y=1 \mid x)}} \approx \beta_j' \Delta x$$

in words: "the odds of j to k increase by 100 $\beta_j' \Delta x$ % if x increases by one unit Δx"

Example: choice by electorate (`faraway`). choice variable:

$$Y = \begin{cases} \texttt{Democrats} & j=1 \\ \texttt{Independent} & j=2 \\ \texttt{Republican} & j=3 \end{cases}$$

model

$$\mathbb{P}\left(Y = y_j \middle|\ \mathbf{x}\right) = \begin{cases} \frac{1}{1+\sum_{j>1}\exp\left(\alpha_j+\beta_j'\mathbf{x}\right)} & \text{for } j=1 \\ \frac{\exp\left(\alpha_j+\beta_j'\mathbf{x}\right)}{1+\sum_{j>1}\exp\left(\alpha_j+\beta_j'\mathbf{x}\right)} & \text{for } j>1 \end{cases}$$

```
multinom(formula = sPID ~age + educ + nincome, data = nes96)
```

Coefficients:

	(Intercept)	age	educ.L	educ.Q	educ.C
Independent	-1.197260	0.0001534525	0.06351451	-0.1217038	0.1119542
Republican	-1.642656	0.0081943691	1.19413345	-1.2292869	0.1544575

	educ^4	educ^5	educ^6	nincome
Independent	-0.07657336	0.1360851	0.15427826	0.01623911
Republican	-0.02827297	-0.1221176	-0.03741389	0.01724679

Interpret as log odds:

$$\log \frac{\mathbb{P}(Y=j \mid \mathbf{x})}{\mathbb{P}(Y=1 \mid \mathbf{x})} = \alpha_j + \mathbf{x}'\beta_j$$

or

$$\log \frac{\mathbb{P}(Y=\texttt{Independent} \mid \mathbf{x})}{\mathbb{P}(Y=\texttt{Democrat} \mid \mathbf{x})} = -\texttt{1.197260} + .. + \texttt{0.01623911} * \texttt{income} + ...$$

ie with every \$1K additional income the log odds of indeps to democs increase by `0.01623911`
(can also be interpreted as %-wise change in odds ceterus paribus, i.e., all other variables in the model)

Grouped data remember: reg on cont regressors $x_1, \dots, x_k$

$$\mathbb{P}\left(Y = j \mid x\right) = \begin{cases} \frac{\exp\left(\alpha_j+\sum_{i=1}^I \beta_{ij}x_i\right)}{1+\sum_{j=2}^J \exp\left(\alpha_j+\sum_{i=1}^I \beta_{ij}x_i\right)} & \text{wenn } \ 1<j\le J \\ \frac{1}{1+\sum_{j=2}^J \exp\left(\alpha_j+\sum_{i=1}^I \beta_{ij}x_i\right)} & \text{wenn } \ j=1 \end{cases}$$

apply: reg on dummy variables. try regressors $x_1, \dots, x_i$ group indicators $I\left(\text{group}\left(1\right)\right), \dots, I\left(\text{group}\left(I\right)\right)$
problem: overparametrization by linear dependence of covariates

$$\sum\nolimits_{i=1}^I I\left(\text{group}\left(i\right)\right) = 1 \implies \alpha_j+\sum\nolimits_{i=1}^I \beta_{ij} I\left(\text{group}\left(i\right)\right) = \alpha_j+\beta_{1j}+\sum\nolimits_{i=1}^I \left(\beta_{ij}-\beta_{1j}\right) I\left(\text{group}\left(i\right)\right)$$

$\Longrightarrow$ ***identifiability conditions***

$$\beta_{1j} = 0 \text{ for all } j = 1, \ldots, J$$

Restrictions:

cont cov $x_1, \ldots, x_I$

	1	$\cdots$	J
x_1	$\alpha_1 + \beta_1 x_1 = 0$	$\alpha_2 + \beta_2 x_1$	$\cdots$
$\vdots$	$\alpha_1 + \beta_1 x_2 = 0$	$\alpha_2 + \beta_2 x_2$	
x_I	$\vdots$		

factor group indicators $X = 1, \ldots, I$

	1	$\cdots$	J
1	$\alpha_1 + \beta_{11} = 0$	$\alpha_2 + (\beta_{12} = 0)$	$\cdots$
$\vdots$	$\alpha_1 + \beta_{21} = 0$	$\alpha_2 + \beta_{22}$	
I	$\vdots$		

model

$$\mathbb{P}(Y = j \mid X = i) = \frac{\exp\left(\alpha_j + \sum_{i=1}^{I} \beta_{ij} I(\text{group}(i))\right)}{\sum_{j=1}^{J} \exp\left(\alpha_j + \sum_{i=1}^{I} \beta_{ij} I(\text{group}(i))\right)}$$
$$\alpha_1 = \beta_{i1} = \beta_{1j} = 0$$

or in short form

$$\log \frac{\mathbb{P}(Y = j \mid X = i)}{\mathbb{P}(Y = 1 \mid X = i)} = \alpha_j + \beta_{ij}$$
$$\alpha_1 = \beta_{i1} = \beta_{1j} = 0$$

or explicitly

$$\mathbb{P}(Y = j \mid X = i) = \begin{cases} \frac{\exp(\alpha_j + \beta_{ij})}{1 + \sum_{j=2}^{J} \exp(\alpha_j + \beta_{ij})} & \text{wenn } 1 < j \leq J \\ \frac{1}{1 + \sum_{j=2}^{J} \exp(\alpha_j + \beta_{ij})} & \text{wenn } j = 1 \end{cases}$$
$$\alpha_1 = \beta_{i1} = \beta_{1j} = 0$$

Note that in the case of binomial regressions:

$$\mathbb{P}(Y = j \mid X = i) = \begin{cases} \frac{\exp(\alpha + \beta_i)}{1 + \exp(\alpha + \beta_i)} & \text{wenn } 1 < j \leq 2 \\ \frac{1}{1 + \exp(\alpha + \beta_i)} & \text{wenn } j = 1 \end{cases}$$

interpretation (multinomial)**:**

remember the matrix of exponents

X \ Y	1	2	$\cdots$	j	$\cdots$	J
1	0	α_2		α_j		α_2
2	0	$\alpha_2+\beta_{22}$				$\alpha_J+\beta_{2J}$
$\vdots$		$\alpha_2+\beta_{32}$				$\alpha_J+\beta_{3J}$
			$\ddots$			$\vdots$
i				$\alpha_j+\beta_{ij}$		
$\vdots$					$\ddots$	
I	0	$\cdots$		$\alpha_j+\beta_{Ij}$		$\alpha_J+\beta_{IJ}$

hence

$$\left.\begin{array}{ll} \log\left(p_{j|i} \,/\, p_{1|i}\right) = \alpha_j + \beta_{ij} & \text{if } \ 2 \le i \le I \\ \log\left(p_{j|1} \,/\, p_{1|1}\right) = \alpha_j & \text{if } \ i = 1 \end{array}\right\}$$

$\Longrightarrow$

$$\boxed{\beta_{ij} = \begin{cases} \log\left(\frac{p_{j|i}}{p_{1|i}} \Big/ \frac{p_{j|1}}{p_{1|1}}\right) & \text{if } \ 2 \le i \le I \text{ and } j > 1 \\ 0 & \text{if } i = 1 \text{ or } j = 1 \end{cases}}$$

and

$$\boxed{\alpha_j = \begin{cases} \log \frac{p_{j|1}}{p_{1|1}} & \text{if } 2 \le j \le J \\ 0 & \text{if } j = 1 \end{cases}}$$

or

$$\frac{\mathbb{P}\left(Y = j \mid X = i\right)}{\mathbb{P}\left(Y = 1 \mid X = i\right)} = \exp \beta_{ij} \frac{\mathbb{P}\left(Y = j \mid X = 1\right)}{\mathbb{P}\left(Y = 1 \mid X = 1\right)}$$

in words:

"the chances of $Y = j$ to $Y = 1$ in the i^{th} group are $\exp \beta_{ij}$ times the chances in the of $Y = j$ to $Y = 1$ in the reference group"

$$\frac{\mathbb{P}\left(Y = j \mid X = 1\right)}{\mathbb{P}\left(Y = 1 \mid X = 1\right)} = \exp \alpha_j$$

in words:

"the chances of $Y = j$ to $Y = 1$ in the reference group"

$$\frac{\mathbb{P}\left(Y = j \mid X = i\right)}{\mathbb{P}\left(Y = 1 \mid X = i\right)} = \exp \alpha_j + \beta_{ij}$$

in words:

"the chances in the of $Y = j$ to $Y = 1$ in the i^{th} group"

4.3.2 ML-estimation

Likelihoods:

case: **person-level** data: n individuals with respective covariates (continuous) $x_1, \ldots, x_n$ and a response vector $(y_1, \ldots, y_J)$ which marks 1 for the category chosen and 0 for all other categories

Data: individuals / categories

	1	$\cdots$	J
x_1	0	$\ldots 1 \ldots$	0
$\vdots$	$\vdots$		

model

$$\mathbb{P}\left(Y = j \mid x\right) = \begin{cases} \frac{\exp(\alpha_j + \beta_j x)}{1 + \sum_{j=2}^{J} \exp(\alpha_j + \beta_j x)} & \text{wenn } \ 1 < j \le J \\ \frac{1}{1 + \sum_{j=2}^{J} \exp(\alpha_j + \beta_j x)} & \text{wenn } \ j = 1 \end{cases}$$

Likelihood (on person level): Define $y_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{else} \end{cases}$ for person i with covariate x_i $\Longrightarrow$

$$L\left(\alpha, \beta \mid \mathbf{y}, \mathbf{x}\right) = \prod_{i=1}^{n} \prod_{j=1}^{J} \mathbb{P}\left(y_i = j \mid x_i\right)^{y_{ij}}$$

$\Longrightarrow$

$$
\begin{aligned}
& l\,(\alpha, \beta \mid \mathbf{y}) \\
&= \sum\nolimits_{i=1}^{n} \left(y_{i1} \log \frac{1}{1 + \sum_{j=2}^{J} \exp\left(\alpha_j + \beta_j x_i\right)} + \sum\nolimits_{j=2}^{J} y_{ij} \log \frac{\exp\left(\alpha_j + \beta_j x_i\right)}{1 + \sum_{j=2}^{J} \exp\left(\alpha_j + \beta_j x_i\right)} \right) \\
&= \sum\nolimits_{i=1}^{n} \left(-y_{i1} \log \left\{ 1 + \sum\nolimits_{j=2}^{J} \exp\left(\alpha_j + \beta_j x_i\right) \right\} \right. \\
&+ \left. \sum\nolimits_{j=2}^{J} -y_{ij} \log \left\{ 1 + \sum\nolimits_{j=2}^{J} \exp\left(\alpha_j + \beta_j x_i\right) \right\} + \sum\nolimits_{j=2}^{J} y_{ij} \log \exp\left(\alpha_j + \beta_j x_i\right) \right) \\
&= \sum\nolimits_{j=2}^{J} \sum\nolimits_{i=1}^{n} y_{ij} \left(\alpha_j + \beta_j x_i\right) - n \log \left(1 + \sum\nolimits_{j=2}^{J} \exp\left(\alpha_j + \beta_j x_i\right) \right)
\end{aligned}
$$

if x_i is a group variable, taking only k different values, this can be decomposed in $k-1$ group indicators(with weights $\beta_{jk} = 0$). the outer sum can be taken of the different values (groups) of x, hence it will be a sum over the number of summands for each i.e. the frequency for each value of x. hence the person-level data and the group level data have the same likelihood in this case.

the mle is obtained by solving the score equations

$$
\begin{aligned}
&\frac{\partial}{\partial \alpha_j} l\,(\alpha, \beta \mid \mathbf{y}) = 0 \\
&\frac{\partial}{\partial \beta_j} l\,(\alpha, \beta \mid \mathbf{y}) = 0, \; j = 1, \ldots, J
\end{aligned}
$$

(start by imposing identifiability conditions)

4.3.3 The Poisson-multinomial connection

one factor multinomial: for categorical data. transform a two factor loglinear model into a one factor multinomial model. identify parameters.

consider log linear model $(X, Y)\,,\; X = x_1, \ldots, x_I,\; Y = y_1, \ldots, y_J$

$$
\begin{aligned}
\mu_{ij} &= \exp\left(\mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij}\right) \\
\alpha_1 &= \beta_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0
\end{aligned}
$$

Equivalently, since $\sum_i \sum_j \mu_{ij}$

$$
\left.
\begin{array}{r}
\mu_{ij} = \mu p_{ij} \\
\mu = \sum_i \sum_j \mu_{ij}
\end{array}
\right\} \Longrightarrow p_{ij} = \frac{\mu_{ij}}{\sum_i \sum_j \mu_{ij}} = \frac{\exp\left(\mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij}\right)}{\sum_i \sum_j \exp\left(\mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij}\right)}
$$

$$
= \frac{\exp\left(\alpha_i + \beta_j + (\alpha\beta)_{ij}\right)}{\sum_i \sum_j \exp\left(\alpha_i + \beta_j + (\alpha\beta)_{ij}\right)}
$$

$$
\text{with} \quad \alpha_1 = \beta_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0
$$

furthermore

$$p_{j\ |i} = \frac{p_{ij}}{p_{i\cdot}} = \frac{p_{ij}}{\sum_j p_{ij}} = \frac{\exp\left(\mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij}\right) \Big/ \sum_i \sum_j}{\sum_j \exp\left(\mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij}\right) \Big/ \sum_i \sum_j}$$

$$= \frac{\exp\left(\beta_j + (\alpha\beta)_{ij}\right)}{\sum_j \exp\left(\beta_j + (\alpha\beta)_{ij}\right)}$$

$$\text{with } \ \beta_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0$$

or, changing notation: $\boxed{\alpha_j \leftrightarrow \beta_j}$

$$\mathbb{P}\,(Y = j \mid X = i) = \begin{cases} \begin{array}{|c|c|} \hline \frac{\exp\left(\alpha_j + (\alpha\beta)_{ij}\right)}{1+\sum_{j>1}\exp\left(\alpha_j+(\alpha\beta)_{ij}\right)} & \text{if } 1 < i \le I \\ \hline \frac{\exp(\alpha_j)}{1+\sum_{j>1}\exp(\alpha_j)} & \text{if } i = 1 \\ \hline \end{array} & \text{if } 1 < j \le J \\ \begin{array}{|c|c|} \hline \frac{1}{1+\sum_{j>1}\exp\left(\alpha_j+(\alpha\beta)_{ij}\right)} & \text{if } 1 < i \le I \\ \hline \frac{1}{1+\sum_{j>1}\exp(\alpha_j)} & \text{if } i = 1 \\ \hline \end{array} & \text{if } \ j = 1 \end{cases}$$

or

$$\frac{\mathbb{P}\,(Y = j \mid X = i)}{\mathbb{P}\,(Y = 1 \mid X = i)} = \exp\left(\alpha_j + (\alpha\beta)_{ij}\right)$$

or

$$\log p_{j\ |i} \big/ p_{1\ |i} = \alpha_j + (\alpha\beta)_{ij}$$

with with $\ \alpha_1 = (\alpha\beta)_{i1} = (\alpha\beta)_{1j} = 0$

In this sense the one factor multinomial and the two factor Poisson models are equivalent. The ML-estimates from a log linear model for grouped data can be identified with the parameters in a multinomial logit for grouped data.

Two factor multinomial: This relation between more general loglinear models and multinomial regression factor models in the analysis of contingency tables is found easily. For example, consider the relation between a three factor loglinear and a two factor multinomial model.

$$\log \mu_{ijk} = \mu_0 + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$$

with side conditions. hence

$$p_{k_0|ij} = \frac{p_{ijk_0}}{p_{\cdot ij}} = \frac{\exp\left(\alpha_j + \beta_i + \gamma_{k_0} + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik_0} + (\beta\gamma)_{jk_0} + (\alpha\beta\gamma)_{ijk_0}\right)}{\sum_k \exp\left(\alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}\right)}$$

$$= \frac{\exp\left(\gamma_{k_0} + (\alpha\gamma)_{ik_0} + (\beta\gamma)_{jk_0} + (\alpha\beta\gamma)_{ijk_0}\right)}{\sum_k \exp\left(\gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}\right)}$$

Note: after renaming this model could be written as

$$\log \mathbb{P}(z=k \mid x,y)/\mathbb{P}(z=1 \mid x,y) = \alpha_z + \beta_{xz} + \gamma_{yz} + (\beta\gamma)_{xyz}$$
$$\alpha_1 = \beta_{1z} = \beta_{x1} = \gamma_{1z} = \gamma_{y1} = 0$$

Example: smoking-survival-data (Simpson's paradox)

survival: $Z = \begin{cases} 1 & \text{dead\_no} \\ 2 & \text{dead\_yes} \end{cases}$ smoking: $X = \begin{cases} 1 & \text{smoke\_no} \\ 2 & \text{smoke\_yes} \end{cases}$

age: $Y = \left\{ \begin{array}{|l|l|l|l|l|} \hline \text{18-25} & \text{25-34} & ... & ... & ... \\ \hline 1 & 2 & 3 & 4 & 5 \\ \hline \end{array} \right.$

model: (no 3-way interaction)

$$\log \mu_{xyz} = \mu_0 + \alpha_x + \beta_y + \gamma_z + (\alpha\beta)_{xy} + (\alpha\gamma)_{xz} + (\beta\gamma)_{yz}$$

with side conditions, or

$$p_{z \mid x,y} = \frac{p_{zxy}}{p_{\cdot xy}} = \frac{\exp\big(\mu_0+\alpha_x+\beta_y+\gamma_z+(\alpha\beta)_{xy}+(\alpha\gamma)_{zx}+(\beta\gamma)_{zy}\big)}{\sum_z \exp\big(\mu_0+\alpha_x+\beta_y+\gamma_z+(\alpha\beta)_{xy}+(\alpha\gamma)_{zx}+(\beta\gamma)_{zy}\big)}$$

canceling and modifying notation:

$$p_{z \mid x,y} = \frac{\exp\Big(\gamma_z + (\gamma\alpha)_{zx} + (\gamma\beta)_{zy}\Big)}{\sum_z \exp\Big(\gamma_z + (\alpha\gamma)_{xz} + (\beta\gamma)_{yz}\Big)}$$
$$\gamma_1 = (\gamma\alpha)_{1x} = (\gamma\alpha)_{z1} = (\gamma\beta)_{1y} = (\gamma\beta)_{z1} = 0$$

or,

$$p_{z \mid x,y} = \begin{cases} \frac{\exp\big(\gamma_z+(\gamma\alpha)_{zx}+(\gamma\beta)_{zy}\big)}{1+\sum_{z>1}\exp\big(\gamma_z+(\alpha\gamma)_{xz}+(\beta\gamma)_{yz}\big)} & \text{for } \ 1 < z \le K \\ \frac{1}{1+\sum_{z>1}\exp\big(\gamma_z+(\alpha\gamma)_{xz}+(\beta\gamma)_{yz}\big)} & \text{for } \ z = 1 \end{cases}$$

hence

$$\log \mathbb{P}(z \mid x,y)/\mathbb{P}(z=1 \mid x,y) = \gamma_z + (\gamma\alpha)_{zx} + (\gamma\beta)_{zy}$$

or

$$\log \frac{\mathbb{P}(z \mid x,y)/\mathbb{P}(z=1 \mid x,y)}{\mathbb{P}(z \mid 1,y)/\mathbb{P}(z=1 \mid 1,y)} = (\gamma\alpha)_{zx}$$

example:

$$\log \frac{\mathbb{P}(\text{dead} \mid \text{smoke}, y)/\mathbb{P}(\text{live} \mid \text{smoke}, y)}{\mathbb{P}(\text{dead} \mid \text{nsmoke}, y)/\mathbb{P}(\text{live} \mid \text{nsmoke}, y)} = (\gamma\alpha)_{\text{dead:smoke}}$$

Estimation results
$(\gamma\alpha)_{\text{dead:smoke}} = \texttt{0.42741}$
or
$\mathbb{P}(\text{dead} \mid \text{smoke})/\mathbb{P}(\text{live} \mid \text{smoke}) = \exp(\texttt{0.42741})\, \mathbb{P}(\text{dead} \mid \text{no\_smoke})/\mathbb{P}(\text{live} \mid \text{no\_smoke})$
i.e.
chances(dead | smoke) are 42% higher than for no_smoke in all age groups
since $\exp(\texttt{0.42741}) \approx 1 + \texttt{0.42741}$

```
glm(formula = y ~ (smoker + dead + age)^2, family = "poisson",
    data = femsmoke)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.10629    0.12759  32.184  < 2e-16 ***
smokeryes           -0.13074    0.18539  -0.705 0.480667
deadyes             -3.86011    0.59389  -6.500 8.05e-11 ***
age25-34             0.92648    0.15071   6.147 7.87e-10 ***
age35-44             0.61204    0.15767   3.882 0.000104 ***
age45-54             0.08972    0.17335   0.518 0.604774
age55-64             0.27942    0.16550   1.688 0.091355 .
age65-74            -0.73124    0.21519  -3.398 0.000678 ***
smokeryes:deadyes    0.42741    0.17703   2.414 0.015762 *
smokeryes:age25-34  -0.11752    0.22091  -0.532 0.594749
smokeryes:age35-44  -0.01268    0.22800  -0.056 0.955654
smokeryes:age45-54   0.56538    0.23585   2.397 0.016522 *
smokeryes:age55-64  -0.08512    0.23573  -0.361 0.718030
smokeryes:age65-74  -1.49088    0.30039  -4.963 6.93e-07 ***
deadyes:age25-34     0.12006    0.68655   0.175 0.861178
deadyes:age35-44     1.34112    0.62857   2.134 0.032874 *
deadyes:age45-54     2.11336    0.61210   3.453 0.000555 ***
deadyes:age55-64     3.18077    0.60057   5.296 1.18e-07 ***
deadyes:age65-74     5.08798    0.61951   8.213  < 2e-16 ***
---
Call:
glm(formula = cbind(y.dead, y.live) ~ smoker + age, family =
binomial(link = logit))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8601     0.5939  -6.500 8.05e-11 ***
smokeryes     0.4274     0.1770   2.414 0.015762 *
age25-34      0.1201     0.6865   0.175 0.861178
age35-44      1.3411     0.6286   2.134 0.032874 *
age45-54      2.1134     0.6121   3.453 0.000555 ***
age55-64      3.1808     0.6006   5.296 1.18e-07 ***
age65-74      5.0880     0.6195   8.213  < 2e-16 ***
```

further details The correspondence among ML-estimators can be explained as follows.

First, remember the relation between independent Poissons and Multinomial: multivariate Poisson conditioned on the total number of observations is a multinomial

$N_1, \ldots, N_I$ independent, $N_i \sim \text{Pois}(\mu_i)\,,\ i = 1, \ldots, I$

$$\begin{aligned}\mathbb{P}\,(n_1, \ldots, n_I \mid \mu_1, \ldots, \mu_I) &= \mathbb{P}\,(n_1, \ldots, n_I \mid p_1, \ldots, p_I, n.)\,\mathbb{P}\,(n. \mid \mu.)\\ &= \text{Mnom}\,(n_1, \ldots, n_I \mid p_1, \ldots, p_I, n.)\,\text{Pois}\,(n. \mid \mu.)\end{aligned}$$

where $p_1, \ldots, p_{I-1}, \mu.$ and $\mu_1, \ldots, \mu_I$ are equivalent sets of parmeters with $p_i = \mu_i \,/\, \mu.$ with $\mu. = \sum_i \mu_i$

Hence, $(\widehat{p}_1, \ldots, \widehat{p}_I) = \arg\max l\,(p_1, \ldots, p_I \mid n_1, \ldots, n_I)$ can be found, by
1. finding $(\widehat{\mu}_1, \ldots, \widehat{\mu}_I) = \arg\max l\,(\mu_1, \ldots, \mu_I \mid n_1, \ldots, n_I)$
2. finding $\widehat{\mu}. = \arg\max l\,(\mu. \mid n.) = \sum_i \widehat{\mu}_i$
3. transforming $(\widehat{p}_1, \ldots, \widehat{p}_I) = (\widehat{\mu}_1 \,/\, \widehat{\mu}., \ldots, \widehat{\mu}_I \,/\, \widehat{\mu}.)$

Similarly
Multivariate Poisson (2-dim contingency table):
$\log(\mu_{ij}) = \mu_0 + \delta_i + \gamma_j + \phi_{ij}$

$\Longrightarrow$ Likelihood:
$L\,(\delta, \gamma, \phi \mid n_{11}, \ldots, n_{IJ}) = \prod_{ij} \mathbb{P}\,(n_{ij} \mid \mu_{ij}) = \prod_{ij} \mu_{ij}^{n_{ij}} \exp(-\mu_{ij})$

$= \prod_i \mathbb{P}\,(n_{i1}, \ldots, n_{iJ} \mid \mu_{i1}, \ldots, \mu_{iJ})$
$= \prod_i \mathbb{P}\,(n_{i1}, \ldots, n_{iJ} \mid \mu_{i1}, \ldots, \mu_{iJ}, n_{i\cdot})\,\mathbb{P}\,(n_{i\cdot} \mid \mu_{i1}, \ldots, \mu_{iJ})$

$= \prod_i \mathbb{P}\,(n_{i1}, \ldots, n_{iJ} \mid \mu_{i1}, \ldots, \mu_{iJ}, n_{i\cdot})\,\mathbb{P}\,(n_{i\cdot} \mid \mu_{i\cdot})$

$= \prod_i \text{Mnom}(y_{i1}, \ldots, y_{iJ} \mid p_{i1}, \ldots, p_{iJ})\text{Pois}(n_{i\cdot} \mid \mu_{i\cdot})$ with $p_{ij} = \mu_{ij}/\mu_{i\cdot}$
where

$$\mu_{i\cdot} = \sum_j \mu_{ij} = \sum_j \exp(\mu_0 + \delta_i + \gamma_j + \phi_{ij}) = \exp(\mu_0 + \delta_i) \sum_j \exp(\gamma_j + \phi_{ij})$$

is the marginal Poisson mean (not a function of j) which carries no information on j.

$$\log \mu_{i\cdot} = \mu_0 + \delta_i^*$$

with

$$\delta_i^* = \delta_i + \log \sum\nolimits_j \exp(\gamma_j + \phi_{ij})$$

$\Longrightarrow$

$$\begin{aligned}\log p_{ij} = \log \mu_{ij} - \log \mu_{i\cdot} &= \mu_0 + \delta_i + \gamma_j + \phi_{ij} - (\mu_0 + \delta_i^*)\\ &= \gamma_j + \phi_{ij} - \log \sum\nolimits_j \exp(\gamma_j + \phi_{ij})\end{aligned}$$

or

$$p_{ij} = \frac{\exp\left(\gamma_j + \phi_{ij}\right)}{\sum_j \exp\left(\gamma_j + \phi_{ij}\right)}$$

Hence the maximizer of the lhs (Poisson likelihood) also maximizes the rhs (multinomial likelihood), hence the mle $\widehat{\phi}_{ij}$ and $\widehat{\gamma}_j$ in $\log\left(\mu_{ij}\right) = \mu_0 + \delta_i + \gamma_j + \phi_{ij}$ equal the mle $\widehat{\beta}_{ij}$ and $\widehat{\alpha}_j$ in $\log\left(p_{ij} \ / \ p_{i1}\right) = \alpha_j + \beta_{ij}$

4.3.4 Multinomial Probit

Utility maximization (latent variable approach)

$$\mathbf{u} = \mathbf{x}' B + \epsilon$$

where

$$\epsilon \sim \mathcal{N}_p\left(\mu, \Sigma\right)$$

coordinatewise

$$u_j = \mathbf{x}'\beta_j + \epsilon_j \ , j = 1, \ldots, J$$

$\Longrightarrow$

$$y = j \Leftrightarrow u_j = \max_{1 \le j \le J} u_j$$

$\Leftrightarrow$

$$\begin{aligned} \mathbb{P}\left(Y = j \mid \mathbf{x}\right) &= \mathbb{P}\left(u_j > u_j \text{ for all } 1 \le j \le J \mid \mathbf{x}\right) \\ &= \mathbb{P}\left(\epsilon_j - \epsilon_j > \mathbf{x}'\left(\beta_j - \beta_j\right) \text{ for all } 1 \le j \le J, \ j \ne j \mid \mathbf{x}\right) \end{aligned}$$

Likelihood

$$L\left(\beta_1, \ldots, \beta_J \ , \Sigma \mid y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n\right) = \prod_i^n \prod_j^J \mathbb{P}\left(y_i = j \mid \mathbf{x}_i, \beta_1, \ldots, \beta_J, \Sigma\right)^{y_{ij}}$$

with $y_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{else} \end{cases}$

Notes:

1. very hard to evaluate this likelihood, only possible for low number of categories, alternatively use MCMC.
2. correlation between the choices on similar alternatives is part of the model
3. for the case $\epsilon_1, \ldots, \epsilon_J$ iid Weibull (Gumbel type 1) with density ("double exponential")

   $$f\left(\epsilon\right) \sim \exp\left(-\exp\left(-\epsilon\right)\right)$$

   the resulting model is the multivariate logit. The independence assumption of disturbances rules out correlation between correlated alternatives (**I**ndependence **I**rrelevant **A**lternatives property)

4.3.5 Multivariate Probit

The probit modell is easily generalized to the multivariate case because the unterlying normal distributional assumptions in the latent variable model are easily generalized to the multinormal case.

This is not so straight forward for the multinomial logit. For example, consider the case that two multinomial observations (maybe from different multinomial distributions and not necessarily independent) are made for each individual.

$$
\begin{aligned}
y^{*(1)} &= \alpha^{(1)} + \beta^{(1)} x + \varepsilon^{(1)} \\
y^{*(2)} &= \alpha^{(2)} + \beta^{(2)} x + \varepsilon^{(2)} \\
\begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{bmatrix} &\backsim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sum \right)
\end{aligned}
$$

Suppose the binary case: $y^{(1)}, y^{(2)} = 0, 1$

$$
\begin{aligned}
\mathbb{P}\left(y^{(1)} = 1, y^{(2)} = 1 \middle| \; x \right) &= 1 - \mathbb{P}\left(y^{(1)} = 1, y^{(2)} = 1 \middle| \; x \right) \\
&= \mathbb{P}\left(\varepsilon^{(1)} > -\left(\alpha^{(1)} + \beta^{(1)} x \right), \; \varepsilon^{(2)} > -\left(\alpha^{(2)} + \beta^{(2)} x \right) \middle| \; x \right) \\
&= \int_{-\alpha^{(1)} - \beta^{(1)} x}^{\infty} \int_{-\alpha^{(2)} - \beta^{(2)} x}^{\infty} \varphi(t_1, t_2) \; dt_1 \; dt_2
\end{aligned}
$$

ParameterSchätzung mit MaxLike.