Could be typos.

15/03/16 DW

1

08/03/16 ALSM2

Multinomial Distribution

Extension to Binomial Distribution

Binomial is a joint probability distribution

$P(y_1, \dots, y_J \mid \theta_1, \dots, \theta_J, n) = \frac{n!}{y_1! y_2! \dots y_J!} \theta_1^{y_1} \theta_2^{y_2} \dots \theta_J^{y_J}$

J: number of categories

For example J=3 yes/no/maybe

y_1: # yes y_2: # no y_3: # maybe

$\theta_1, \dots, \theta_J$ are the respective probabilities of the categories $\theta_1 + \dots + \theta_J = 1$

$n = y_1 + y_2 + \dots + y_J$

$E[y_1] = n\theta_1 \quad E[y_2] = n\theta_2 \dots$

- Main difference so far → represent variable in a vector, not a scale

Exercise 1: Multinomial with J=2

$P(y_1, y_2 \mid \theta_1, \theta_2, n) = \frac{n!}{y_1! y_2!} \theta_1^{y_1} \theta_2^{y_2} \quad n = y_1 + y_2 \quad \theta_1 + \theta_2 = 1$

$P(y_1 \mid \theta_1, n) = \frac{n!}{y_1!(n-y_1)!} \theta_1^{y_1} (1-\theta_1)^{n-y_1}$ Binomial Distribution

Exercise 2: Member of exponential family of distributions?

J = 2 yes, shown for binomial dist.

J > 2 → NO

- Even if multinomial dist is not a member of the exponential family of distributions, we can control $\vec{\theta}\vec{y}$ collected over N groups via a set of parameters $\vec{\beta}$

Nominal Logistic Regression

Definition: The outcomes of experiments are in J categories and there IS no natural order amongst the response categories. One category is arbitrarily chosen as a reference category eg θ_1 then the logits for the other categories are defined by:

$$\text{Logit}(\theta_j) = \left(\frac{\theta_j}{\theta_1}\right) = x^T\beta_j \quad \forall j = 2, \dots, J$$

having the constraint $\sum_{j=1}^{J} \theta_j = 1$ when the estimates $\hat{\beta}_j$ are computed, then

$$\hat{\theta}_j = \hat{\theta}_1 \exp(x^T\hat{\beta}_j) \quad \forall j = 2, \dots, J$$

$$\hat{\theta}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp(x^T\hat{\beta}_j)}$$

or $\hat{\theta}_j = \dfrac{\exp(x^T\hat{\beta}_j)}{1 + \sum_{j=2}^{J} \exp(x^T\hat{\beta}_j)} \quad \forall j = 2, \dots, J$

Softmax function → converting output to probability

Can rewrite as $\hat{\theta}_1 + \hat{\theta}_1 \exp(x^T\hat{\beta}_2) + \dots + \hat{\theta}_1 \exp(x^T\hat{\beta}_J) = 1$

$$\hat{\theta}_1 = \frac{1}{1 + \exp(x^T\hat{\beta}_2) + \dots + \exp(x^T\hat{\beta}_J)}$$

Example From Website

- Split into 6 groups
- Calculate θ_{ij} for each possible: 6 groups 3 variables

Model 1: $\text{Log}\left(\frac{\theta_j}{\theta_1}\right)^{j=2,3} = \beta_{0j} + \beta_{1j}\text{Sex} + \beta_{2j}\text{Age} + \beta_{4j}\text{Age}^2$

Model 2: $\text{Log}\left(\frac{\theta_j}{\theta_1}\right) = \beta_{0j} + \beta_{1j}\text{Sex} + \beta_{2j}\text{Age}$

Model 4 has two less β's hence lower AIC

Model 2 has reduced complexity, not entirely a better fit

2

Could be typos. 15/03/16 DW

08/03/16 ALSM 2 3

R output: (Model 1)

$y \sim Sex + age1 + age^2$

	intercept	Sex	age1	age2
y_2	-0.59	-0.34	1.13	1.59
y_3	-1.04	-0.81	1.48	2.92

What are the $\hat{\theta}$s for model 1?

Use softmax function, $\hat{\theta}_j = \frac{\exp(x^T\beta_j)}{1 + \sum_{j=2}^{J} \exp(x^T\beta_j)}$ $\forall j = 2, J$

	y_1	y_2	y_3
$i =$ 1		(12)	(13)
2			
3			
4			
5			
6			

(12) $\rightarrow j=2, i=1 = \frac{\exp(-0.59)}{1 + \exp(-0.59) + \exp(-1.04)} = 0.29$

(13) $\rightarrow j=3, i=1 = \frac{\exp(-1.04)}{1 + \exp(-0.59) + \exp(-1.04)}$

For $\hat{\theta}_2$ (the overall y_2 value, NOT per individual group)

$\hat{\theta}_{2i} = \frac{\exp(-0.59 - 0.34(Sex_i) + 1.13(Age1) + 1.59(Age2))}{1 + \exp(-0.59 - 0.34Sex + 1.13(Age1) + 1.59(Age2)) + \exp(-1.04 - 0.81(Sex) + 1.48 Age1 + 2.92 Age2)}$ group i, category j

Same for θ_3 $\hat{\theta}$ is $1 - \hat{\theta}_2 - \hat{\theta}_3$

ODDS RATIO

$\frac{\theta_{21}/\theta_{11}}{\theta_{24}/\theta_{14}} = 1.5$ — women important / women not important (18-21) / men important / men not important 18-21

odds ratio assumes use of Softmax function

- Proportion of women of important over not important is bigger than that of men who voice important over not important for age group 18-23

Multinomial Distribution

- Instead of value of response, have a vector of responses (y)
- Observe a number of groups with vector responses and # people in each group

$\{(\vec{y_i}, x_i, n_i)\}_{i=1,\dots,N}$ (number of groups)

- n_i: # people voting in group i.
- x_i: Explanatory variable associated with group i.

J categories, J-1 d.f. $\#\beta(J-1) = D.F.$

i.e. intercept, Sex, age $= 3 \times 2 = 6$ D.F. in example

D.F of Saturated model 2 d.f for each group (3 options per group)

6 groups = 12 D.F. for Saturated model

When AIC differ by 2 between models, could be from an extra variable ($2m$) in AIC formula, m is $m+1$, hence extra 2

Comparing M_1 and M_2

- M_1 8 parameters, M_2 - 6
- Difference between log likelihoods $-2[\log L_{M_1} - \log L_{M_2}] \approx 2$ (596 - 594)
- Are the models different? $2 \sim \chi^2_2$
- 2 will be in the 95% CI of chi-Square
- We can say $M_1 \approx M_2$, basically equivalent.

Could be typos.

15/08/16 DW

08/03/16 ALSM 2

ODDS RATIO i = group j = vote choice

$OR = \frac{\theta_{j=2,i=1}}{\theta_{j=1,i=1}}$ ratio for group 1 women (0-2) Sex=0 Age1=0 Age2=0

$\theta_{j=2,i=4} / \theta_{j=1,i=4}$ ratio for group 2 men (1-2) Sex=1 age1=0 age2=0

Category 2 = Y_2 $j=1$ = reference category

Only explanatory variable changing between groups is sex.

OR = 1.5 ⇒ indicates the group is more important to women than men (used log as link fn)

$$\text{Log}\left[\frac{\theta_{j=2,i=1}}{\theta_{j=1,i=1}}\right] = \log\left[\frac{-0.59 - 0.38\,\text{Sex}_i + 1.24\,\text{Age1}_i + 1.58\,\text{Age2}_i}{-0.59 - 0.38\,\text{Sex}_i + 1.24\,\text{Age1}_i + 1.58\,\text{Age2}_i}\right]$$

Compute for $i=1$ and $i=4$ and compute ratio

Only thing that will change is the $-0.38\,\text{Sex}_i$

How to use S.E. to ensure odds ratio is correct?

OR for Sex = $\exp[+0.8 \pm 2*0.3]$ 0.3 is S.E. for Sex variable

2*S.E. of Sex ⇒ Sex was being tested with 3

CI $[(\exp(0.38 - 2*0.3), (\exp(0.38 + 2*0.3))]$ = $[0.9, 2.68]$ 95% C.I.

- because OR related on only one explanatory variable
- 1 is within interval ⇒ men and women may be the same

- Compare group 3 to 6 because Age2 and Age1 will change ⇒ new result
- Should provide same result as previous test and end up with same β
- Can compare 3 to category 1 (use β's), 1 will be inside the interval "not important" chosen as reference category

Could be typos. 15/03/16 DW

15/03/16 ALSM2

ALLIGATORS EXAMPLE - MULTINOMIAL

A. -Multinomial distribution is an extension of the binomial distribution where the number of categories possible for the is larger than 2
-Here the number of categories is J=5, corresponding to 5 food choices the alligator

$$f(y_1,y_2,y_3,y_4,y_5|\theta_1,\theta_2,\theta_3,\theta_4,\theta_5) = \frac{n}{y_1!y_2!y_3!y_4!y_5!}\theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3}\theta_4^{y_4}\theta_5^{y_5}$$

with $y_1+y_2+y_3+y_4+y_5=n$ and $\theta_1+\theta_2+\theta_3+\theta_4+\theta_5=1$

B. When we know 3 pieces of information we will know 4th piece

Hancock: $(L_1,L_2,L_3) = (1,0,0)$
Oklawaha: $(L_1,L_2,L_3) = (0,1,0)$
Trafford: $(L_1,L_2,L_3) = (0,0,1)$
George: $(L_1,L_2,L_3) = (0,0,0)$

C $\text{Log}\left[\frac{\theta_j}{\theta_1}\right] = \beta_j^{\text{intercept}} + \beta_{j,\text{sex}} + \beta_{j,\text{size}} + \beta_{j,L1} + \beta_{j,L2} + \beta_{j,L3} \quad \forall j=2,\dots,5$

$$\theta_j = \frac{\exp(\beta_j^T x)}{1+\sum_{j=2}^{J}(\beta_j^T x)} \quad \forall j=2,\dots,5 \text{ with } x=[1,\text{Sex},\text{Size},L_1,L_2,L_3]$$

$\beta_j = [\beta_j^{\text{intercept}}, \beta_j^{\text{sex}}, \beta_j^{\text{size}}, \beta_j^{L1}, \beta_j^{L2}, \beta_j^{L3}]$

$$\theta_1 = \frac{1}{1+\sum_{j=2}^{5}\exp(\beta_j^T x)}$$

D $L = \prod_{i=1}^{16} P(y_{1i},y_{2i},y_{3i},y_{4i},y_{5i}\,|\,\theta_{1i},\theta_{2i},\theta_{3i},\theta_{4i},\theta_{5i})$

$n_i = y_{1i}+y_{2i}+y_{3i}+y_{4i}+y_{5i}$ group sizes can be different

$$\theta_{1i} = \frac{1}{1+\sum_{j=2}^{J}\exp(\beta_j^T x)}$$ ← used to calculate fitted values in table 8.

$$\theta_{ji} = \frac{\exp(\beta_j^T x_i)}{1+\sum_{j=2}^{J}\exp(\beta_j^T x_i)}$$

$L = \prod_{i=1}^{16} \frac{n_i}{y_{1i}! y_{2i}! y_{3i}! y_{4i}! y_{5i}!} \theta_{1i}^{y_{1i}} \theta_{2i}^{y_{2i}} \theta_{3i}^{y_{3i}} \theta_{4i}^{y_{4i}} \theta_{5i}^{y_{5i}}$ $df = 4 \times 16 - 45$

Interval

$L(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ $df = 6 \times 4$

$L(\{\beta_j\}, j = 2, \ldots, 5)$

E AIC - criterion to select best model. Can use to drop parameter or not

F. $\beta_2^T X = 0.17 - 0.468x - 1.345 size - 1.796_1 + 0.916_2 + 1.166_3$

$\beta_3^T X = -3.42 - 0.638x + 0.756 size + 1.134 + 2.553_2 + 3.062_3$

$\beta_4^T X = -2.43 - 0.815x + 0.735 size + 0.384_1 - 0.556_2 + 1.246_3$

$\beta_5^T X = -1.43 - 0.255x - 0.245 size + 0.774 + 0.226_3 + 1.566_4$

G. Use table of fitted vote (table 8

Hancock	vs	Oklahoma	OOD RATIO
i=1 $\frac{0.05}{0.6} = 0.83$		i=5 $0.03/0.3 = 0.026$	3.08
i=2 $0.11/0.32$		i=6 $0.028/0.43$	3.08
i=3 $0.074/0.51$		i=7 $0.04/0.21$	3.08
i=4 $0.17/0.52$		i=8 $0.038/0.36$	3.08

OR = $\exp(\beta_{j=4}^{L1} - \beta_{j=4}^{L2}) = 1.13$

H Confident?

S.E. $\beta_{j=6}^{L1} = 0.79$ $\beta_{j=6}^{L2} = 1.2$

take: $SE(\beta_{j=6}^{L1} - \beta_{j=6}^{L2})^2 = S.E(\beta_{j=6}^{L1})^2 + S.E(\beta_{j=6}^{L2})^2$ independence assumed.

$S.E = \sqrt{0.79^2 + 1.2^2} = 1.45$

$1.13 \pm 1.45 \times 2 = 95\%$ CI

0 is within the interval $\Rightarrow$ not confident

$\exp(0) = 1$, there is a chance that OR can or will be 1.

Can't say for sure but "allycat) in ..... , than in late..."

Could be types.

15/08/16 DW

4. Alligators

The table 7 comes from a study of the primary food choices of alligators in four Florida lakes. Researchers classified the stomach contents of 219 captured alligators into five categories: Fish (the most common primary food choice), Invertebrate (snails, insects, crayfish, etc.), Reptile (turtles, alligators), Bird, and Other (amphibians, plants, household pets, stones, and other debris).

i	Lake	Sex	Size	Fish y_1	Inver. y_2	Rept. y_3	Bird y_4	Other y_5
1	Hancock	M	small	7	1	0	0	5
2			large	4	0	0	1	2
3		F	small	16	3	2	2	3
4			large	3	0	1	2	3
5	Oklawaha	M	small	2	2	0	0	1
6			large	13	7	6	0	0
7		F	small	3	9	1	0	2
8			large	0	1	0	1	0
9	Trafford	M	small	3	7	1	0	1
10			large	8	6	6	3	5
11		F	small	2	4	1	1	4
12			large	0	1	0	0	0
13	George	M	small	13	10	0	2	2
14			large	9	0	0	1	2
15		F	small	3	9	1	0	1
16			large	8	1	0	0	1

Table 7: Alligators primary Food Choice.

The expert decides to use the multinomial distribution to model the responses $\{y_1, \cdots, y_5\}$.

(a) Explain why this distribution is suited for this problem.

(b) The expert defines the following explanatory variables:

$$\text{sex} = \begin{cases} 0 & \text{female} \\ 1 & \text{male} \end{cases} \quad \text{size} = \begin{cases} 0 & \text{small} \\ 1 & \text{large} \end{cases} \quad \text{Lake}_1 = \begin{cases} 1 & \text{If Hancock} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Lake}_2 = \begin{cases} 1 & \text{If Oklawaha} \\ 0 & \text{otherwise} \end{cases} \quad \text{Lake}_3 = \begin{cases} 1 & \text{If Trafford} \\ 0 & \text{otherwise} \end{cases}$$

Explain why only 3 indicator variables are defined to encode the information about the 4 possible lakes where the data has been collected?

(c) Assuming the reference category is θ_1 (proportion for *fish*), write down the linear model linking the proportions $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ with the explanatory variables.

(d) Explain what function is used and maximised to compute the parameters of the models (noted βs in the course). Write down explicitly this function with the notation you have introduced in question 4c.

(e) The model is fitted using R and the output in R is:

```
Call:
multinom(formula = y.mat ~ sex + size + Lake1 + Lake2 + Lake3)
```

```
Coefficients:
    (Intercept)        sex       size      Lake1      Lake2     Lake3
y2   0.1690702 -0.4629388 -1.3361658 -1.7805555  0.91304120 1.155722
y3  -3.4161432 -0.6276217  0.5571846  1.1296426  2.53024945 3.061087
y4  -2.4321397 -0.6064035  0.7300740  0.5754599 -0.55020075 1.237216
y5  -1.4309095 -0.2524299 -0.2905697  0.7667093 -0.02603021 1.557820

Residual Deviance: 537.8655
AIC: 585.8655
```

Explain the meaning of the AIC and its use.

(f) Rewrite the model find in question 4c using the numerical values found with R.

(g) The fitted values are also computed automatically with R and these are reported in table 8.

i	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$
1	0.6006304	0.07545645	0.032585176	0.051157366	0.24017062
2	0.6236286	0.02059329	0.059063749	0.110228530	0.18648582
3	0.5070764	0.10120786	0.051529843	0.079201241	0.26098463
4	0.5157553	0.02705796	0.091497934	0.167174242	0.19851457
5	0.3034156	0.56356125	0.066792314	0.008384288	0.05784657
6	0.4825231	0.23557649	0.185433818	0.027670279	0.06879629
7	0.2146236	0.63333355	0.088499072	0.010875857	0.05266792
8	0.3591723	0.27859182	0.258550999	0.037770774	0.06591415
9	0.2088132	0.49437446	0.078160321	0.034470791	0.18418127
10	0.3050677	0.18984756	0.199345737	0.104509741	0.20122923
11	0.1449154	0.54508513	0.101605098	0.043869783	0.16452463
12	0.2132216	0.21081022	0.260984348	0.133952044	0.18103180
13	0.5008607	0.37332877	0.008780762	0.023993546	0.09303619
14	0.6826618	0.13374903	0.020893106	0.067865670	0.09483043
15	0.3930845	0.46549198	0.012908441	0.034531949	0.09398311
16	0.5781329	0.17995524	0.033143452	0.105397407	0.10337101

Table 8: Estimated fitted values.

The expert says: *alligators in Lake Oklawaha are less likely to choose birds over fish than their colleagues in Lake Hancock are.* Explain mathematically where this conclusion comes from.

(h) Assuming the βs independent, how confident are you in this conclusion by the expert given the standard errors of the parameters computed by R:

```
Std. Errors:
    (Intercept)       sex      size     Lake1     Lake2     Lake3
y2   0.3787475 0.3955162 0.4111827 0.6232075 0.4761068 0.4927795
y3   1.0851582 0.6852750 0.6466092 1.1928075 1.1221413 1.1297557
y4   0.7706720 0.6888385 0.6522657 0.7952303 1.2098680 0.8661052
y5   0.5381162 0.4663546 0.4599317 0.5685673 0.7777958 0.6256868
```

(25 marks)