ETH zürich

Applied Generalized Linear Models (FS 20)

Nominal outcomes: multinomial logistic regression

Viviana Amati

Course structure - schedule

Date	Topic	Assignment
18.02	Introduction to the course	Ass. 1 released on 25.02 due to 19.03
25.02	Introduction to R and review of the linear regression mode	
03.03	The general linear model: ANOVA and ANCOVA	
10.03	Practical: ANOVA and ANCOVA	
17.03	Binary outcomes: logistic regression and probit models	Ass. 2 released on 18.03 due to 23.04
24.03	**Practical: logistic regression and probit models**	
31.03	**Nominal outcomes: multinomial logistic regression**	
07.04	Practical: multinomial logistic regression	
21.04	Ordinal outcomes: ordered logistic regression and probit models	Ass. 3 released on 25.04 due to 21.05
28.04	Practical: ordered logistic regression and probit models	
05.05	Count outcomes: Poisson and negative binomial models	
12.05	Practical: Poisson and negative binomial models	
19.05	Survival models	L+P
26.05	Panel data model	L+P

Exam

1. Assignment 4
   4 exercises for each student, to be returned in one week

2. Analyse an assigned data set
   Results should be presented in a report (max. 3000 words)

3. Deepen a topic that we have not treated (extension of what we learned)
   Report (max. 6 pages) including a short example

Today's agenda

- Logistic regression analysis practical
- Introduction to multinomial logistic regression
- Lecture and slides

Logistic regression: Data

The data set `admission.csv` contains information on the admission of 400 students into a business school. The variables in the data set are:

- *admit*: binary variable taking value 1 if the student was admitted into the business school and 0 otherwise
- *gpa*: grade point average in the undergraduate institution (range 1 – 6)
- *gre*: graduate record examination score obtained in the undergraduate institution (range 0 – 1000)
- *rank*: prestige of the undergraduate institution. The variable takes on the values 1 (highest prestige) through 4 (lowest prestige).

Test the association between admit and all the other variables

Logistic regression model

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_{\text{gre}} X_{\text{gre}} + \beta_{\text{gpa}} X_{\text{gpa}} + \beta_{\text{r2}} D_{\text{r2}} + \beta_{\text{r3}} D_{\text{r3}} + \beta_{\text{r4}} D_{\text{r4}}$$

with D_r the dummy variables for rank with reference category highest prestige (1).

Logistic regression model

- Odds ratio

$$OR = \frac{\pi(\mathbf{x}+1)/[1-\pi(\mathbf{x}+1)]}{\pi(\mathbf{x})/[1-\pi(\mathbf{x})]} = e^{\beta_j} \quad .$$

  OR=1 no association between Y and X_j.

- Wald-type confidence interval (CI) at level $\alpha = 0.05$

$$\left[e^{\beta_j - 1.96 \times s.e.(\beta_j)}, e^{\beta_j + 1.96 \times s.e.(\beta_j)}\right]$$

  If the confidence interval includes 1, the OR is not significantly different from 1

- Profile CI: robust wr.t. small sample size and asymmetries

Grouped data

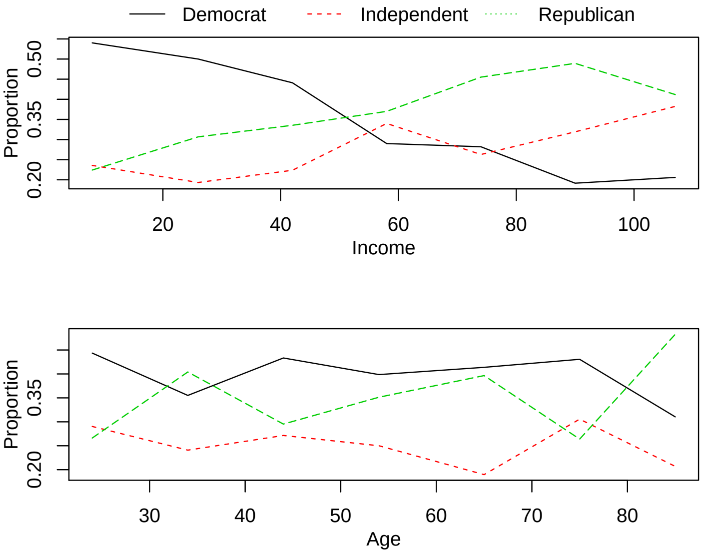
The titanic.csv data set

Economic status	Age group	Gender	Survived	Died	Total
Crew	A	W	20	3	23
Crew	A	M	192	670	862
1st	A	W	140	4	144
1st	A	M	57	118	175
2nd	A	W	80	13	93
2nd	A	M	14	154	168
3rd	A	W	76	89	165
3rd	A	M	75	387	462
1st	C	W	1	0	1
1st	C	M	5	0	5
2nd	C	W	13	0	13
2nd	C	M	11	0	11
3rd	C	W	14	17	31
3rd	C	M	13	35	48
Total			711	1490	2201

An example

- Data from the 1996 American National Election Study
  (Rosenstone, Kinder, and Miller (1997))
- Information on
  - Party identification of the respondent
    (Democrat, Independent or Republican)
  - age
  - income (thousand of dollars)
- Is there an association between the party identification and the other variables?

An example


Democrat
Independent
Republican
Proportion
0.20
0.35
0.50
20
40
60
80
100
Income
Proportion
0.20
0.35
30
40
50
60
70
80
Age

Multinomial logistic regression model

- Nominal dependent variable with $M > 2$ categories
  (categories do not have a natural order)

- Simultaneously use all pairs of categories by specifying the odds of success in one category instead of another

$$\log\left[\frac{\pi_a(\mathbf{x})}{\pi_b(\mathbf{x})}\right], \quad a, b \in \{0, 1, \ldots, M\}, a \neq b$$

$$\pi_a(\mathbf{x}) = P(Y = a | \, \mathbf{X}) \quad \pi_b(\mathbf{x}) = P(Y = b | \, \mathbf{X})$$

- Pairing each category with the reference category M is enough to describe all the log-odds

$$\text{logit}[\pi_m(\mathbf{x})] = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p$$

(baseline-category logits)

Multinomial logistic regression model (MNRM)

$$\text{logit}[\pi_m(\mathbf{x})] = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p$$

Multinomial logistic regression model (MNRM)

$$\text{logit}[\pi_m(\mathbf{x})] = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p$$

- Each of the $M-1$ logits has its own parameter

  The model can have a large number of parameters

Multinomial logistic regression model (MNRM)

$$\text{logit}[\pi_m(\mathbf{x})] = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p$$

- Each of the $M-1$ logits has its own parameter

  The model can have a large number of parameters

- The $M-1$ log-odds are enough to describe all the $\binom{M}{2}$ pairs of categories

$$\begin{aligned}\log\left[\frac{\pi_a(\mathbf{x})}{\pi_b(\mathbf{x})}\right] &= \log\left[\frac{\pi_a(\mathbf{x})/\pi_M(\mathbf{x})}{\pi_b(\mathbf{x})/\pi_M(\mathbf{x})}\right] = \log\left[\frac{\pi_a(\mathbf{x})}{\pi_M(\mathbf{x})}\right] - \log\left[\frac{\pi_b(\mathbf{x})}{\pi_M(\mathbf{x})}\right] \\ &= (\beta_{0a}-\beta_{0b}) + (\beta_{1a}-\beta_{1b})X_1 + \ldots + (\beta_{pa}-\beta_{pb})X_p\end{aligned}$$

Probabilities

- For $m = 1, \dots, M-1$

$$\pi_m(\mathbf{x}) = \frac{\exp(\beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p)}{1 + \sum_{h=1}^{M-1} \exp(\beta_{0h} + \beta_{1h}X_1 + \ldots + \beta_{ph}X_p)}$$

- For the reference category

$$\pi_M(\mathbf{x}) = 1 - \sum_{m=1}^{M-1} \pi_m(\mathbf{x})$$

MNRM as a multivariate GLM

The MNRM is a multivariate GLM where:

- the random component $Y|\mathbf{X}$ has a multinomial distribution with

$$\pi_m(\mathbf{x}_i) = \mathrm{E}[Y_i = m|\mathbf{X}]$$

- the link function g_m for each category is the logit

$$g_m(\mathrm{E}[Y = m|\mathbf{X}]) = \log\left[\frac{\pi_m(\mathbf{x})}{\pi_M(\mathbf{x})}\right]$$

- the systematic component for each category is

$$\eta_m = \sum_{j=1}^{p} \beta_{jm} X_j$$

Estimation

- Based on MLE
- No close form for the solution

  Newton-Raphson algorithm or its variants
- The ML estimator is MVUE and has asymptotic distribution

$$\mathbf{B} \sim N\left(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta})\right)$$