PRACTICE 7: POLYTOMOUS OUTCOME

## R software possibilities:

* Nominal response: nnet library multinom(Y ~ A+B+C) any formula is valid, where Y is a polytomous response variable and A, B and C are factors (qualitative variables) .
* Be careful with the default order of factor levels :.
  + Reorder to simplify interpretation: factor(variable, levels=c(nivell1, …, nivellsk))
  + If factor levels are not meaningful include labels for factor levels: factor(variable, levels=c(nivell1, …, nivellsk),labels=c(nom1,…,nomk)).
* Logarithm link function by default is log(probability\_i/ probability\_ref.).
* Ordinal response: polr(Y ~ A+B+C) any formula is valid, where Y is a polytomous response variable and A, B and C are factors (qualitative variables). A latent variant approach is considered in R.
* Hierarchical logit response. First level work vs nonwork and Second level parttime vs fulltime.

## CASE 1: Female labor integration (Fox)

The generalized linear model that is proposed investigates the analysis of the relationship between young married women who work based on the existence of children in the home, the income of their husbands and the region of the country where they reside (WOMENLF file).

• The response variable is polytomous and has 3 categories: does not work (1), works part time (2) and works full time (3). The baseline category is not working.

• The presence of children in the home is factor A, which has 2 categories (YES, NO). Base category: NO (the constant corresponds to the mean value of the NO category).

• The Canada region is a polytomous factor B, with 5 categories.

• The husband's income (in thousands of dollars) is covariate X.

• Intuition indicates an interaction between the income of husbands (X) and the presence of children (A).

* Analyze the data using nominal categories.
* Analyze the data assuming ordinal categories.
* Analyze the data using a hierarchical model.

### Analyze data using nominal response models (multinom method (.) In the nnet package, neural networks)

1. First of all do a basic exploratory analysis of the data. Note that the covariate has atypical values: write down which observations are affected. Construct a variable that groups income ranges (the covariate), for example using medians as class representatives and gathering the rank observations of the highest covariate (keep the factor and covariant version):

> intervals

[1] 0 5 10 15 20 25 30 46

**womenlf$gincome** <- factor(cut(womenlf$income,breaks=intervals))

table(womenlf$gincome)

cintervals<- tapply(womenlf$income,womenlf$gincome,median)

womenlf$ngincome<-factor(cut(womenlf$income,breaks=intervals))

womenlf$ngincome<-factor(womenlf$ngincome,labels=cintervals)

**womenlf$ngincome**<-as.numeric(levels(womenlf$ngincome))[womenlf$ngincome]

1. Define the binary variable bwork, grouping of the original answer (work), where you classify the observations into work (parttime + fulltime) and not work (not\_work). Fill the table below which leads to choosing by inference criteria as the best A + X model (at home, it is not the aim of this session).

**womenlf$bwork<-as.factor(ifelse(womenlf$work=="not\_work","not\_work","work" ))**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | ***Análisis de la Devianza*** | | | | | |
| ***Model***  **bwork** | | *p* | ***Deviance*** | *ΔDeviance* | *d.f.* | *Comments* | |
| *Contrast* | *P value* |
| **0** | **1** |  |  |  |  |  |  |
| **1** | **A** |  |  |  |  |  |  |
| **2** | **X** |  |  |  |  |  |  |
| **3** | **A+X** |  |  |  |  |  |  |
| **4** | **B** |  |  |  |  |  |  |
| **5** | **A+B** |  |  |  |  |  |  |
| **6** | **A\*X** |  |  |  |  |  |  |
| **7** | **A+B+X** |  |  |  |  |  |  |
| **8** | **B+A\*X** |  |  |  |  |  |  |

1. Determine which of the explanatory variables is the most individually contributing to the explicability of the nominal polytomous response (work) and write the equations that characterize the model:

a. (mm1) A. Write down the deviation explained: G1a.

b. (mm4) B. Write down the deviation explained: G1b.

c. (mm3) A+X. Write the deviation explained: G1c.

d. Choose the best model from the 3 listed above.

1. Determine the best option for modelling nominal polytomous response (work) and write the equations.
2. (Optional) Represent in ordinates and scale defined by the link function the observed data and the value of the superimposed linear predictor in the chosen model. On abscissas the original covariate (Income): identify the points according to the factor A. You will have 2 graphs because there are 2 equations to characterize the model: use aggregation procedure agregacioPOL in terms of classes of the covariate, that is, as many classes as defined by the model A + X (for simplicity).
3. Is it worth adding factor B to the model (A)? Construct the corresponding deviation table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Model*** | | *n-p* | Deviance |  | *Contrast* | *g.l.* | ***Equations for each logodd*** |
| **1** | **1** | 18 |  |  |  |  | Constant: |
| **2** | **A** | 16 |  |  | (M2) vs (M1) |  |  |
| **3** | **B** | 10 |  |  | (M3) vs (M1) |  |  |
| **4** | **A+B** | 8 |  |  | (M4) vs (M2) |  | Additiu: |
|  | (M4) vs (M3) |  |
| **5** | **A\*B** | 0 |  |  | (M5) vs (M4) |  | Interaccions Factors: |

1. Is it worth adding to the chosen model (A), the covariate X? In what way? Construct the corresponding deviation table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Model*** | | *n-p* | **Deviance** |  | *Contrast* | *g.l.* | ***Equation for each logodd*** |
| **1** | **1** |  |  |  |  |  |  |
| **2** | **X** |  |  |  | (M2) vs (M1) |  |  |
| **3** | **A+X** |  |  |  | (M3) vs (M1) |  |  |
| **4** | **A\*X** |  |  |  | (M4) vs (M3) | 1 |  |

1. Assess the need to incorporate nonlinear terms of the covariate and calculate the confusion table for your best nominal model.
2. (Optional) Interpret and diagnose the model taking as a grouping in classes of the covariate the one defined by the resulting model A + X (not A + B + X). Represent in ordinates and scale defined by the link function the observed data and the value of the overlapping linear predictor in the chosen model. On abscissas the covariate (Income): identify the points according to the factor A. You will have 2 graphs because there are 2 equations to characterize the model: as classes of the covariate, use those defined by the model A + X.
3. Interpret the coefficients estimated in the equations: in terms of the predictor and the scale of the odds.
4. Determine the probability predicted by the work model in the 3 categories in the case of households with children and the couple's income on average.
5. Calculate 95% confidence intervals for the variation in odds (there are 2) when moving from childless homes to child homes.
6. Calculate 95% confidence intervals for the variation in odds (there are 2) for each unit (in thousands of dollars) of the couple's income increase (INCOME).
7. Determine the confusion table for your best nominal model and assess prediction capacity.
8. (Optional) Use an aggregation of the covariate as defined in point 1. Repeat the calculation of the A + NGX model:

a. Calculate the columns of the logodds observed for each class of the covariate: there are 2, one fulltime vs not\_work and the other parttime vs not\_work.

b. What conclusions did you draw from it?

1. Predict using manual interpretation and predict() method probability for each polytomous purpose according to the best multinomial model for a household having sons and income on the sample mean. Assess model fit to data using an inferential test.

### Analyze data using hierarchical response models (two nested binary models)

**Analyze data using hierarchical, binary models at each level. The top hierarchy (HL1) faces not\_work versus work and the bottom hierarchy (HL2) faces within working women fulltime vs parttime.**

1. Determine by inference which is the best model for the top hierarchy (binary response). If you prefer, justify the A + X model: nothing needs to be added and nothing can be taken away from it.
2. Diagnose the A + X model at the HL1 level (with default graphs). Study the usual Goodness of Fit stats (Pearson and Deviance) - rate them. Does the model have good predictive skills? Justify the answer in statistical terms.
   1. Interpret the estimated coefficients in the equations: in terms of the predictor and the scale of the odds.
   2. Predict the probability of working/not working for a household having sons ‘present’ and mean income.
3. Determine by inference which is the best model for the lower hierarchy (binary response). If you prefer, justify the A + X model: nothing needs to be added and nothing can be taken away from it.
4. Diagnose model A + X at HL2 level (with default graphs). Study the statistics of Goodness of Fit (Pearson and Deviance): rate them. Justify the answer in statistical terms.
   1. Interpret the estimated coefficients in the equations: in terms of the predictor and the scale of the odds.
   2. Predict the probability of partialtime/fulltime working for a household having sons ‘present’ and mean income.
5. Determine the probability predicted by the work model in the 3 categories under Hierarchical Modelling in the case of households with children and couple's income on average.
6. Without considering an income grouping, which model do you think is more appropriate: the multinomial proposal or the hierarchical proposal? Justify the answer statistically, in addition to graphically interpreting the best model and assessing confusion table.

### Analyze data using ordinal response models (polr() method)

**Analyze the data using ordinal response models with logit link, (MASS polr (.) Method). The order of the answer MUST be not\_work, parttime and fulltime.**

1. Determine which of the explanatory variables is most individually contributing to the explicability of the data and write the equations that characterize the model:
2. (M1a) A. Write down the deviation explained: G1a.
3. (M1b) B. Write down the deviation explained: G1b.
4. (M1c) X. Write the deviation explained: G1c.
5. Choose from the 3 above. Verify that you use the order requested in the answer.
6. Is it worth adding factor B to the chosen model (A)? Construct the corresponding deviation table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Model*** | | *n-p* | **Deviança** |  | *Contrast* | *g.l.* | ***Equacions del Model*** |
| **1** | **1** | 18 |  |  |  |  |  |
| **2** | **A** | 16 |  |  | (M2) vs (M1) |  |  |
| **3** | **B** | 10 |  |  | (M3) vs (M1) |  |  |
| **4** | **A+B** | 8 |  |  | (M4) vs (M2) |  |  |
|  | (M4) vs (M3) |  |
| **5** | **A\*B** | 0 |  |  | (M5) vs (M4) |  |  |

1. Is it worth adding to the chosen model (A), the covariate X? In what way? Construct the corresponding deviation table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Model*** | | *n-p* | ***Deviance*** |  | *Contrast* | *d.f.* | ***Model Equations*** |
| **1** | **1** |  |  |  |  |  |  |
| **2** | **X** |  |  |  | (M2) vs (M1) |  |  |
| **3** | **A+X** |  |  |  | (M3) vs (M1) |  |  |
| **4** | **A\*X** |  |  |  | (M4) vs (M3) | 1 |  |

1. Assess the need to incorporate non-linear terms of the covariate.
2. Interpret and diagnose the model (A + X).
3. Interpret the coefficients estimated in the equations: in terms of the predictor and the scale of the accumulated odds. Interpret the cut points.
4. Interpret the estimated coefficients in the equations: in terms of latent variable.
5. Determine the probability predicted by the work model in the 3 categories in the case of households with children and the couple's income on average.
6. What conclusions did you draw from it? Rate Goodness of Fit stats.
7. Which model do you think is more appropriate: the multinomial proposal or the hierarchical proposal or the ordinal proposal? Statistically justify the answer, in addition to assessing the interpretability of the models.
8. Would gompit or probit models improve model properties?

### Case 2: Housing Conditions in Copenhagen (Madsen-76, Agresti-90)

Data related to the level of satisfaction (low, medium, high - reference low) with the home according to the factor A -Housing -Dwelling type (tower, apartment, atrium, terrace - reference i = 1 tower), the factor C -Influence - Sensation of Influence in the management of the neighborhood community (low, medium, high - reference j = 1 low) and the Factor D –Contact- which is the degree of contact with other residents (low, high - reference k = 1 low). N = 1681.

1. Nominal response. Use additive model A+ C+D using 14 degrees of freedom.

* Model assessment.
* Interprete estimates for level ‘high’ in factor Contact.
* Determine model probability on the reference group for each response category.

1. Ordinal response. Use additive model D+A\*C. Use links: LOGIT, PROBIT and GOMPIT:

* Interprete cut-points.
* Interprete estimates for level ‘high’ in factor Contact.
* Interprete estimates for level ‘high’ in factor Influence.
* Determine model probability for satisfaction on the reference group for each.
* Determine model probability for satisfaction on the reference group for Influence and Housing, but High Contact.
* Select the best link function to be applied.

#### Basic procedure using R

options(contrasts=c("contr.treatment","contr.treatment"))

#ptions(contrasts = c("contr.treatment", "contr.poly"))

copen <- read.table("copenhagen.txt",header=TRUE, sep="\t", na.strings="?")

save.image("W:/seccio\_fme/teaching/mlgz/dades/copenhagen.RData")

summary(copen)

# Ordenar les categories

copen$housing <- ordered(copen$housing, levels=c("tower","apartments","atrium","terraced"))

copen$influence <- ordered(copen$influence, levels=c("low","medium","high"))

copen$contact <- ordered(copen$contact, levels=c("low","high"))

copen$satisfaction <- ordered(copen$satisfaction, levels=c("low","medium","high"))

summary(copen)

attach(copen)

# Model de resposta nominal:

# satisfaccio = housing + influence + contact

library(MASS)

library(nnet)

copen.mult <- multinom(satisfaction~housing+influence+contact,data=copen,weights=n)

summary(copen.mult)

tapply(copen$n,list(copen$satisfaction,predict(copen.mult)),sum)

# on hem obtingut els coeficients dels models per la satisfacció:

# logit 1: medium vs low

# logit 2: high vs low

#Coefficients:

# (Intercept) housingapartments housingatrium housingterraced

#medium -0.4192316 -0.4356851 0.1313663 -0.6665728

#high -0.1387453 -0.7356261 -0.4079808 -1.4123333

# influencemedium influencehigh contacthigh

#medium 0.4464003 0.6649367 0.3608513

#high 0.7348626 1.6126294 0.4818236

# Model de resposta ordinal:

# satisfaccio = housing \* influence + contact

library(MASS)

copen.polr <- polr(satisfaction~housing\*influence+contact,data=copen,weights=n)

summary(copen.polr)

copen.polr$anova

copen.polrp <- polr(satisfaction~housing\*influence+contact,data=copen,weights=n, method="probit" )

summary(copen.polrp)

copen.polrg <- polr(satisfaction~housing\*influence+contact,data=copen,weights=n, method="cloglog")

summary(copen.polrg)

# Questio: Quin tractament sembla mes adequat?

# resposta ordinal? Amb quin dels enllaços

# resposta nominal?

# Respondrem comparant les respostes observades i les esperades

# Això no funciona per ara: cal multiplicar pel tamany de cada classe.

table(copen$satisfaction,predict(copen.polr,satisfaction))

table(copen$satisfaction,predict(copen.mult))

tapply(copen$n,list(copen$satisfaction,predict(copen.mult)),sum)

# Mirem AICs

# Ordinal: AIC: 3484.64 (logistic) AIC: 3485.331 (probit) AIC: 3494.878 (cloglog)

# Multinomial: AIC: 3498.084

**Case 3: Use of Contraception in El Salvador (FESAL-1985) (\*)**

We want to study the response "Use of Contraception" according to "Age", from the perspective of the conditional distribution of the response, Y, given the predictor, "Age". The answer has 3 categories 'Ster', 'Others' and 'None' and the age is grouped in blocks of 5 years between 15 and 50. The contingency table associated with the data (joint distribution analysis perspective) it is:

MTB > Table 'Age' 'Contra\_method';

Rows: Age Columns: Contra\_m

None Other Ster All

15-19 78,38 20,61 1,01 100,00

232 61 3 296

156,28 45,73 93,99 296,00

20-24 64,83 22,20 12,97 100,00

400 137 80 617

325,75 95,33 195,92 617,00

25-29 46,45 20,22 33,33 100,00

301 131 216 648

342,12 100,12 205,76 648,00

30-34 37,11 13,89 48,99 100,00

203 76 268 547

288,80 84,51 173,69 547,00

35-39 43,22 11,49 45,29 100,00

188 50 197 435

229,66 67,21 138,13 435,00

40-44 48,52 7,10 44,38 100,00

164 24 150 338

178,45 52,22 107,33 338,00

45-49 64,44 3,52 32,04 100,00

183 10 91 284

149,94 43,88 90,18 284,00

All 1671 489 1005 3165

Chi-Square = 430,028; DF = 12; P-Value = 0,000

* Analyze the data, determining the adequacy of a linear model. The initial prospecting of the data for the fit of a nominal multinomial model through the representation of the observed probabilities (or log-odds of 'Ster' versus 'None' and of 'Others' vs. 'None') in ordered versus the midpoint of age groups in abscissa suggests that logits are quadratic or cubic functions of 'Age'.
* Address a nominal approach and select the best model.
* Address an ordinal approach and select the best model.
* Analyze the data using a hierarchical model: HL1 'Some' vs. 'None', HL2 'Ster' vs. 'Other'. Retain the best model for each hierarchical level.
* Select the best modelling option for your data.

### Case 4: Cheese tasting (McCullagh)

The study wants to determine the effect on the taste of cheeses, ordinal variable with 9 categories 1 to 9 from significant extremely dislike (1) to excellent taste (9), based on 4 types of additives, ABC and D (base) .

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Additive*** | **Outcome (Y)** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **Total** |
| A | *0* | *0* | *1* | *7* | *8* | *8* | *19* | *8* | *1* | 52 |
| B | *6* | *9* | *12* | *11* | *7* | *6* | *1* | *0* | *0* | 52 |
| C | *1* | *1* | *6* | *8* | *23* | *7* | *5* | *1* | *0* | 52 |
| D | *0* | *0* | *0* | *1* | *3* | *7* | *14* | *16* | *11* | 52 |
| **Total** | 7 | 10 | 19 | 27 | 41 | 28 | 39 | 25 | 12 | 208 |

* Does additive factor affect cheese tasting?
* Interpret Additive factor estimates using lògit transformation.
* Select the most suitable link function to be applied out of logit, probit and gompit.
* Calculate tasting rate probabilities for the reference level in Additive factor.
* Calculate tasting rate probabilities for the D level in Additive factor.