Multinomial logits - model for categorical outcomes

Silje Synnøve Lyder Hermansen

2026-04-09

GLM: A recap

Reminder: What is a GLM?

Regressions aim to describe (a linear) relationship between x and y with one number, β.

- Assumes a continuous and unbounded variable.
- When y is neither (e.g. binary), we relied on a latent continuous variable
- To approximate the latent variable, we calculated the logodds (i.e. we compare)

$\Rightarrow$ *Probability distribution maps unobserved variable to observed outcomes.*

Topic (next weeks): nominal and ordinal variables

GLMs when our outcome variable is categorical

- categorical (e.g. party, profession,. . .) $\rightarrow$ *multinomial or conditional regression*
- ordinal (e.g. attitudes towards topics. . .) $\rightarrow$ *ordinal regression*

Dependent variable: nominal

The discrete choice models describe mutually exclusive choices.

- The choice variable is nominal: we cannot rank it
- Our *appreciation* of it is continuous. Two sets of models:
  - Multinomial: Models *chooser* characteristics
  - Conditional logit: Models *choice* characteristics

Today: multinomial models

- what are they?
- estimation
- interpretation

Warmup: Discuss with your neighbour

Our outcome variable is unordered categorical: it reflects a choice

- How would you model it?
  - how to think about this "choice"?
  - with the models you have already seen?
  - . . . or or we haven't seen?

Two conceptions of multinomial regression

Two conceptions of multinomial regression

- **Latent variable approach:** Our utility of each choice.
- **A series of binomial logits** with the same reference category.

Take 1: Latent variable approach

Intuition

A chooser-centered choice

- each option is more or less attractive:
  - they can be ordered according a latent preference
- each value combination of the predictor is a synthetic "chooser"
  - different choosers prefer different options
  - each chooser trait is given a weight (regression coefficient)

Latent variable approach

Imagine m choice options modeled as $y_m = a_m \times b_m x_i$

- $b_m x_i$ reflects the utility of a choice m for the chooser i with x characteristic. $\rightarrow$ systematic term
- a_m reflects the baseline utility of that choice $\rightarrow$ stochastic term

$\Rightarrow$ *The preferred choice is the one with the highest utility*

Example: Party choice

Let's consider party choice among voters

- ESS survey round (chap 6, Hermansen, 2023)
- respondents give:
  - preferred party
  - attitudes towards immigration

I can rank parties

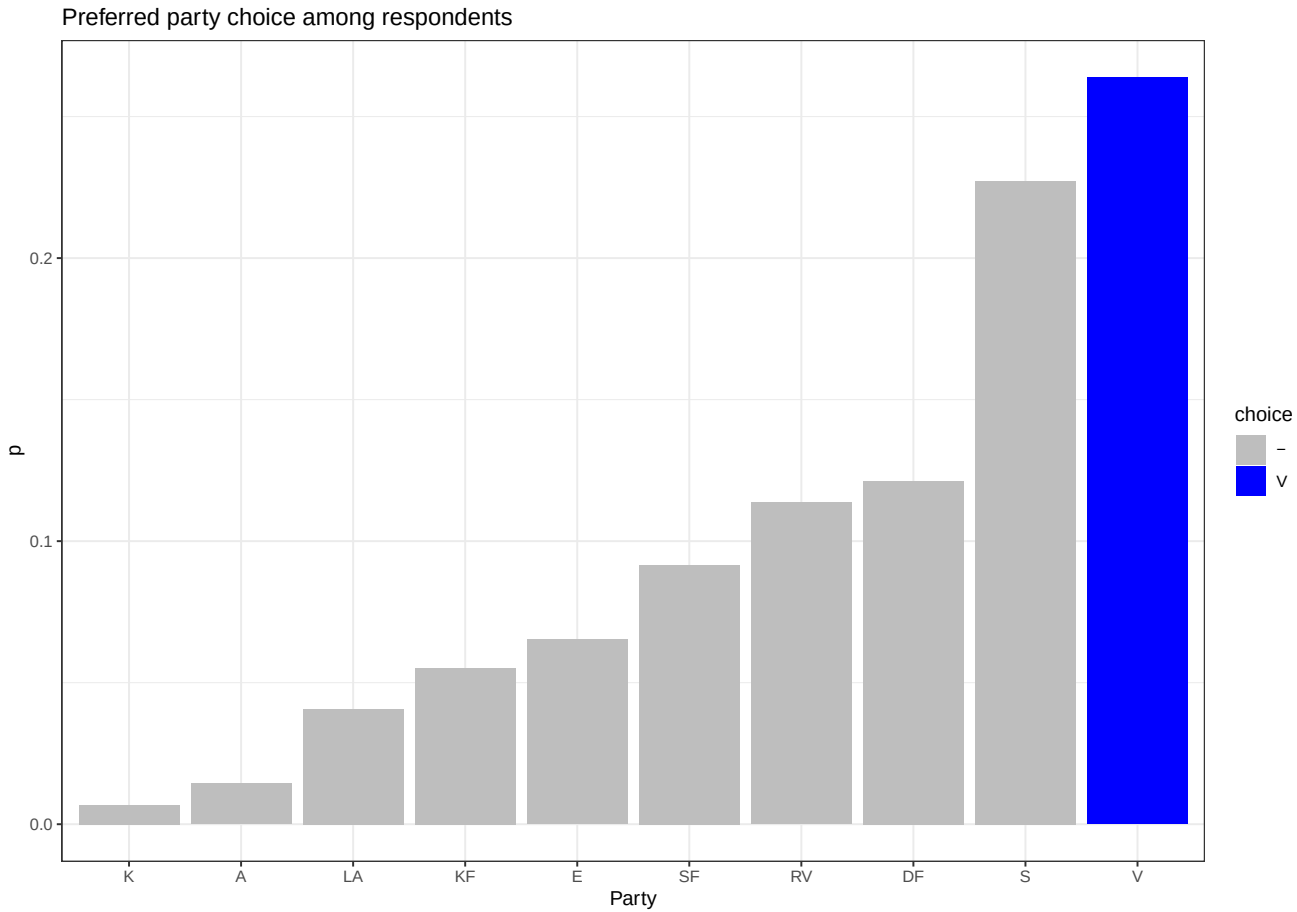
Let's rank the parties according to the respondents' choice

```
tab <-
  df %>%
  #Group by party
  group_by(Party) %>%
  #Number of respondent by party
  reframe(n = n()) %>%
  mutate(
    #Total number of respondents
    N = sum(n),
    #Proportion/probability of group
    p = n/N) %>%
  #Sort just for facility
  arrange(p) %>%
  mutate(
    #Check if it sums up to 1
    cum = cumsum(p),
    #Which is the largest?
    choice = if_else(row.names(.) == which.max(p),
                     Party, "-"))
```

I can rank parties

```
## # A tibble: 10 x 6
##    Party     n     N       p     cum choice
##    <chr> <int> <int>   <dbl>   <dbl> <chr>
##  1 K         8  1179 0.00679 0.00679 -
##  2 A        17  1179 0.0144  0.0212  -
##  3 LA       48  1179 0.0407  0.0619  -
##  4 KF       65  1179 0.0551  0.117   -
##  5 E        77  1179 0.0653  0.182   -
##  6 SF      108  1179 0.0916  0.274   -
##  7 RV      134  1179 0.114   0.388   -
##  8 DF      143  1179 0.121   0.509   -
##  9 S       268  1179 0.227   0.736   -
## 10 V       311  1179 0.264   1       V
```

I can rank parties (figure)



- the most frequent party choice is the most probable outcome

Theoretical link to political science

The assumption is that choosers are rational, and choose a category (m_j) whenever its utility exceeds the alternative (m_d).

$$U(m_j) > U(m_d)$$

$\Rightarrow$ *This is also how we estimate it; through comparisons*

Take 2: A series of binomial logits

Take 2: A series of binomial logits

A series of binomial logits with the *same* reference category.

- Data consists of many groups, but I only compare two groups $\rightarrow$ data/variation intensive model choice.
- Categories/choice are mutually exclusive $\rightarrow$ Different β for each choice

$\Rightarrow$ *All choices are given a probability and they sum up to one.*

Example: ESS survey round

Example: ESS survey round

Let's do an intercept-only model

Logit transformation:

$$logit(p_m) = log(\frac{p_m}{p_d})$$

```
tab <-
  df %>%
  #Group by party
  group_by(Party) %>%
  #Number of respondent by party
  reframe(n = n()) %>%
  mutate(
    #Total number of respondents
    N = sum(n),
    #Proportion/probability of group
    p = n/N,
    #Pick Social democrats as reference category
    p_ref = p[Party == "S"],
    #Odds
    odds = p/p_ref,
    #Logodds
    logodds = log(odds)) %>%
  arrange(logodds)
```

Example: ESS survey

- intercept-only model
- . . . where the reference-level (S) is effectively left out

```
## # A tibble: 10 x 7
##    Party     n     N       p p_ref   odds logodds
##    <chr> <int> <int>   <dbl> <dbl>  <dbl>   <dbl>
##  1 K         8  1179 0.00679 0.227 0.0299  -3.51
##  2 A        17  1179 0.0144  0.227 0.0634  -2.76
##  3 LA       48  1179 0.0407  0.227 0.179   -1.72
##  4 KF       65  1179 0.0551  0.227 0.243   -1.42
##  5 E        77  1179 0.0653  0.227 0.287   -1.25
##  6 SF      108  1179 0.0916  0.227 0.403   -0.909
##  7 RV      134  1179 0.114   0.227 0.5     -0.693
##  8 DF      143  1179 0.121   0.227 0.534   -0.628
##  9 S       268  1179 0.227   0.227 1        0
## 10 V       311  1179 0.264   0.227 1.16     0.149
```

Estimation

Preprosessing: set a reference level

Preprosessing: set a reference level

- We set a reference level p_d: That's the leave-one-out trick.

```
df <-
  df %>%
  #I use the Social democrats
  mutate(Party = relevel(as.factor(Party), ref = "S"))
```

Estimate the model

Null model: no predictors

```
library(nnet)
mod.cat <- multinom(Party ~
                      1,
                    df)
```

```
## # weights:  20 (9 variable)
## initial  value 2714.747825
## iter  10 value 2332.511892
## final  value 2326.831829
## converged
```

Results table

The result is a series of equations, one for each party

Table 1:

	Dependent variable:						
	A	DF	E	K	KF	LA	RV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	-2.76^{***}	-0.63^{***}	-1.25^{***}	-3.51^{***}	-1.42^{***}	-1.72^{***}	-0.69^{*}
	(0.25)	(0.10)	(0.13)	(0.36)	(0.14)	(0.16)	(0.11
Akaike Inf. Crit.	4,671.66	4,671.66	4,671.66	4,671.66	4,671.66	4,671.66	4,671.

Note:

With predictors

Let's regress party choice on scepticism towards immigration

```
library(nnet)
mod.cat <- multinom(Party ~
                      Skepsis,
                    df)
```

```
## # weights:  30 (18 variable)
## initial  value 2705.537484
## iter  10 value 2304.290245
## iter  20 value 2246.392642
## final  value 2246.301290
## converged
```

Table 2:

	Dependent variable:						
	A	DF	E	K	KF	LA	RV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Skepsis	0.23	0.56***	−0.04	−0.18	0.04	0.07	−0.30***
	(0.15)	(0.07)	(0.08)	(0.24)	(0.09)	(0.10)	(0.07)
Constant	−3.89***	−3.69***	−1.04**	−2.70**	−1.58***	−2.06***	0.62*
	(0.85)	(0.40)	(0.41)	(1.09)	(0.44)	(0.51)	(0.32)
Akaike Inf. Crit.	4,528.60	4,528.60	4,528.60	4,528.60	4,528.60	4,528.60	4,528.60

Note: * p

Interpretation

Interpretation

All the possibilities of the binomial logit are open, but the backtransformation is a hack

$\Rightarrow$ *However, you want to decide which story you want to tell*

Different approaches

- With respect to the reference category
  - the regression table (logodds): direction and statistical significance
  - marginal effects (partial back-transformation): relative change
- Predicted outcomes per category
  - predicted probability of each category (transformation of latent variable): when one increases, the other decrease
  - predicted choice (total back-transformation): most probable choice

Marginal effects

The marginal effects are interpreted with reference to the reference level:

- A one-unit increase in skepticism decreases the probability of voting Alternativet rather than Social democrats with:
  - (1 - exp(0.23) $\times$100) = -25%

Predicted probabilities

The results can be read as a series of equations, one for each category m

$$Pr(y = m) \sim log(odds)$$
$$log(odds) = a_m + b_m x$$

- predictions for each category $\rightarrow$ *separate slopes and intercept*

$$log(odds) = -3.89 + 0.23x$$

$\Rightarrow$ *The "latent" variable is here represented by the logodds*

Predicted probabilities (cont.)

The manual backtransformation requires more manual work

1. set a scenario (e.g. x = 5)
2. backtransform: divide the odds for the relevant category by the sum of the odds for all categories (incl. the reference) within each scenario

- calculate the logodds by hand for all categories within the scenario, sum over and exponentiate
- ... or use the `predict()` function in R

```
preds <- predict(mod.cat, newdata = data.frame(Skepsis = 5), type = "probs")
preds
```

```
##           S           A          DF           E           K          KF
## 0.238097927 0.015061927 0.096815641 0.067125356 0.006603907 0.058599993
##          LA          RV          SF           V
## 0.043437070 0.100558519 0.093078122 0.280621538
```

⇒ *The probability that a respondent with moderate view on immigration votes Alternativet is 2 %*

Predicted probabilities using R

Predictions give latent probability of voting for a party, given the scenario.

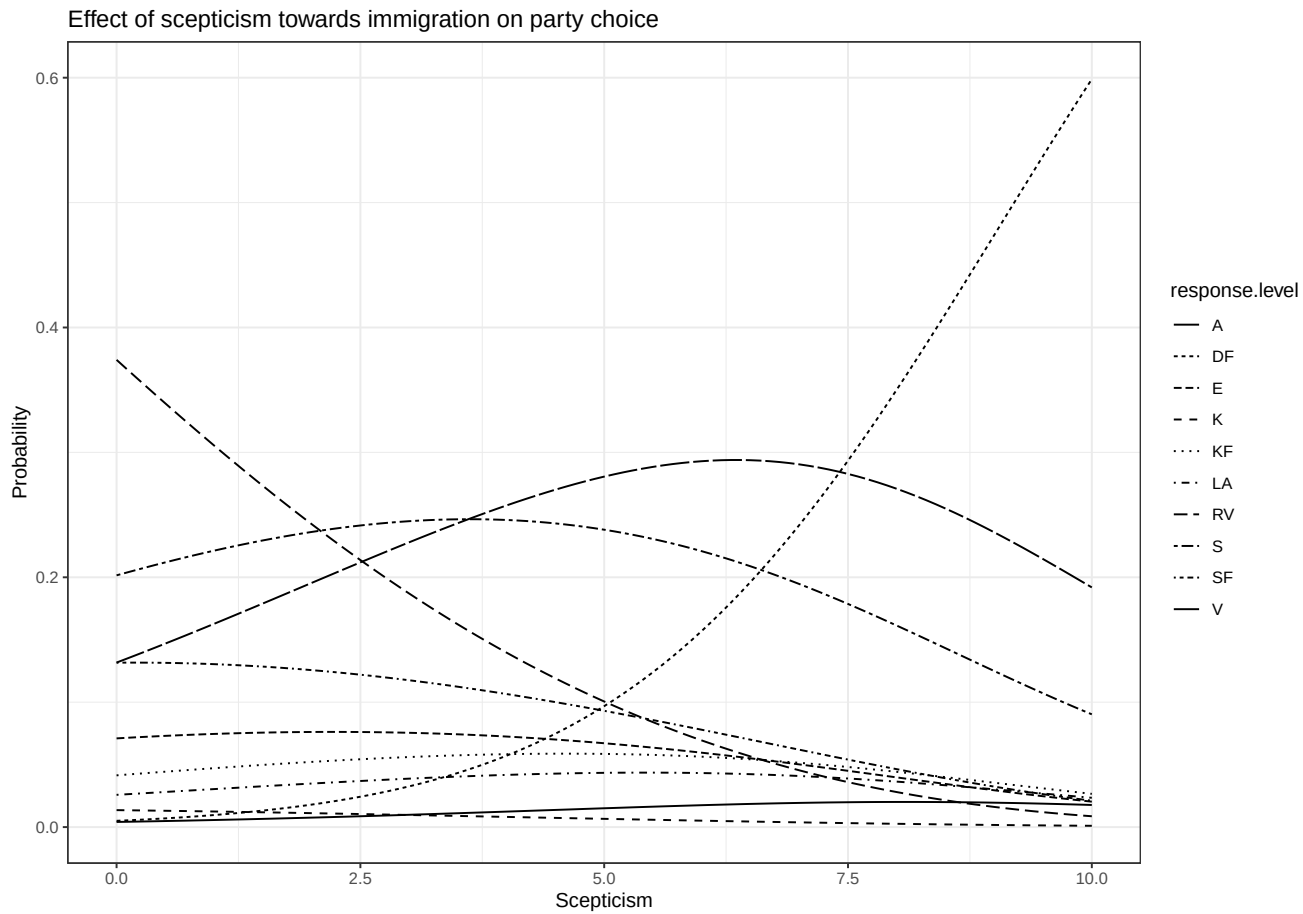
- quickly many predictions

```
predict(mod.cat, newdata = data.frame(Skepsis = 0:10), type = "probs") %>% round(., 3)
```

```
##         S     A    DF     E     K    KF    LA    RV    SF     V
## 1   0.202 0.004 0.005 0.071 0.014 0.041 0.026 0.374 0.132 0.132
## 2   0.221 0.006 0.010 0.075 0.012 0.047 0.030 0.306 0.130 0.163
## 3   0.236 0.008 0.018 0.076 0.011 0.052 0.035 0.243 0.126 0.196
## 4   0.245 0.010 0.033 0.075 0.010 0.056 0.039 0.187 0.118 0.228
## 5   0.246 0.012 0.057 0.072 0.008 0.058 0.042 0.140 0.107 0.258
## 6   0.238 0.015 0.097 0.067 0.007 0.059 0.043 0.101 0.093 0.281
## 7   0.221 0.018 0.157 0.060 0.005 0.056 0.043 0.069 0.078 0.293
## 8   0.195 0.019 0.242 0.050 0.004 0.051 0.041 0.045 0.062 0.290
## 9   0.161 0.020 0.350 0.040 0.003 0.044 0.036 0.028 0.046 0.271
## 10  0.125 0.020 0.474 0.029 0.002 0.036 0.030 0.016 0.032 0.236
## 11  0.090 0.018 0.599 0.020 0.001 0.027 0.023 0.009 0.021 0.192
```

Predicted probabilities using R (cont.)

- in each scenario the sum of probabilities is one:
  - when the probability of voting for one party increases, the probability decreases for other parties
  - lines of effect plot become dependent



Total backtransformation

To predict party choice, I identify the party with the highest probability within each scenario/respondent

- I let the scenario vary (or I can do in-sample prediction) and predict probabilities

```
preds <- predict(mod.cat, newdata = data.frame(Skepsis = 0:10), type = "probs")
```

- I identify the most likely outcome for scenario 1

```
#First scenario
which.max(preds[1,])
```

```
## RV
##  8
```

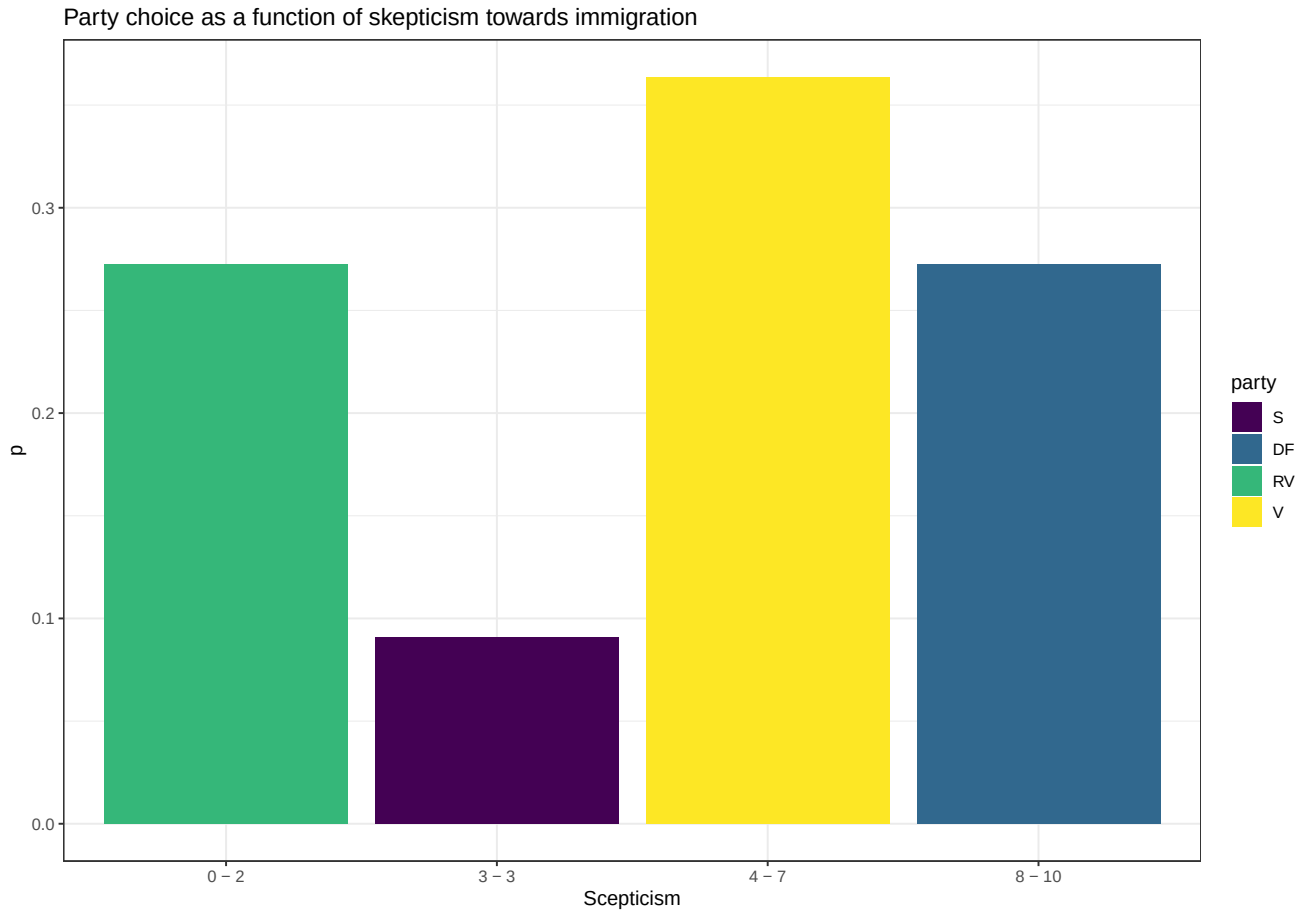
- R can also do it for me

```
preds <- predict(mod.cat, newdata = data.frame(Skepsis = 0:10), type = "class")
preds
```

```
##  [1] RV RV RV S  V  V  V  V  DF DF DF
## Levels: S A DF E K KF LA RV SF V
```

Total backtransformation (cont.)

Predicted categorical outcomes can also be illustrated in a barplot, with the predictor on the x-axis



Main assumption: IIA

Independence of irrelevant alternatives:

- there are no choices beyond what is modeled
- consistency: if we prefer A $>$ B and B $>$ C, then also A $>$ C

$\Rightarrow$ *The β does not depend on on other values of y (other alternatives).*

Testing the main assumption:

The Hausmann-McFadden test: Removes an alternative (supposed to be irrelevant) and check if β changes.

- Restricted model (a choice is removed) vs. unrestricted model (original)
- if IIA holds, then unrestricted model has smaller variance.

$\Rightarrow$ *χ^2-test with smaller value indicates IIA holds.*

Model statistics: Based on predictions

Model statistics: Based on predictions

- **Predict outcome** in sample
  - predicted outcome/choice is the one with the highest probability/utility
  - confusion matrix (Proportion of correct predictions: $\frac{sum\ of\ diagonal}{N\ observations}$)
- **Probability of all outcomes separately**: ROC curve and separation plots

$\Rightarrow$ *as in binomial regression, where you have one category vs. the rest*

Main takeaways

Conceptual

Two ways of understanding multinomial models:

- latent variable approach: maximize utility (latent) when choosing (category)
  - predictors: **chooser-level** characteristics
- estimation: a series of binomial logistic regressions
  - common reference category
  - separate regression coefficients for each choice

Practical

- interpretation:
  - marginal effects: not very intuitive
  - predicted outcomes: probabilities or categories
- IIA assumption:
  - excluding a category from the data, should not change effects