Multinomial and ordered logits

Silje Synnøve Lyder Hermansen

18-11-2019

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 categorical, we rely on a latent continuous variable
- To approximate the latent variable, we calculate the logodds (i.e. we compare)

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

Ordered logistic regression

What is an ordered variable?

A ranked variable with unknown distance between categories.

- Often the result of binning: Close connection to latent formulation.
- We can choose how to treat it: As linear, categorical or **ordinal**.

⇒ *estimate a single set of regression parameters, but keep the information on the order without assuming a continuous variable.*

Two conceptions of ordered logisitc regression

There are two ways of uncerstanding the ordered logit:

- Latent variable: useful for interpretation.
- Parallel regressions: useful for understanding and checking estimation.

Latent variable approach: cutpoints

Cutpoints

We rely on cutpoints to slice up the latent variable and determine outcomes

- **Binomial logistic:** One cutpoint. $\rightarrow$ Rarely estimated.
- **Ordinal logistic:** Serveral cutpoints. $\rightarrow$ Explicit.

$\Rightarrow$ *Model estimates both regression parameters (β) and cutpoints (τ).*

A series of cutpoints

You are in the category m **when the latent variable is between its two cutpoints:** $\tau_{m-1} < y^{\star} < \tau_m$

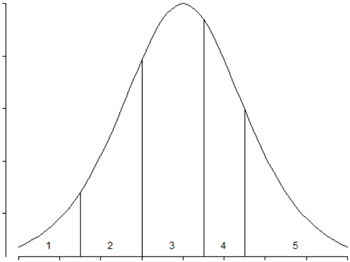


Figure 1: Slicing up a latent variable

The regression coefficients

The model calculates the odds of being lower than τ_m

- The first cutpoint (τ_0) is 0 ($-inf$): you cant be lower than the lowest.
- The last cutpoint is 1 ($+inf$): all observations are in some category.
- You end up with $m-1$ cutpoints.

The regression output

The regression output reports both β and τ

- **Regression coefficient** β is reported in relation to *upper* cutpoint of the category: $\tau_m - \beta x_i$
- **Cutpoints** serve also as intercepts.

The predicted value

The predicted probability of being in category m:

$$Pr(y_i = m) = \frac{exp(\tau_m - \beta x_i)}{1 + exp(\tau_m - \beta x_i)} - \frac{exp(\tau_{m-1} - \beta x_i)}{1 + exp(\tau_{m-1} - \beta x_i)} \tag{1}$$

An example: Attitudes towards redistribution

An example:

ESS respondents (who voted H or FrP) are asked to what extent they believe the state should engage in redistribution (1 = disagree; 5 = agree).

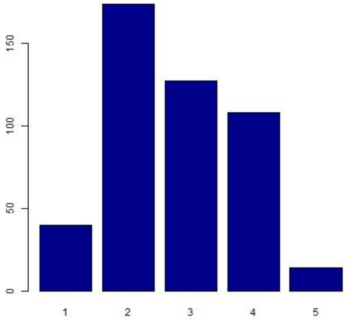


Figure 2: Attitudes towards redistribution is an ordered variable

Attitudes towards redistribution as a function of income

```
## Call:
## polr(formula = Utjevn.ord ~ Inntekt, data = df, Hess = TRUE,
##     method = "logistic")
##
## Coefficients:
##           Value Std. Error t value
## Inntekt 0.08387    0.03128   2.681
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -2.0119  0.2422    -8.3052
## 2|3  0.3029  0.1994     1.5190
## 3|4  1.4724  0.2107     6.9883
## 4|5  3.9305  0.3317    11.8496
##
## Residual Deviance: 1218.94
## AIC: 1228.94
## (17 observations deleted due to missingness)
```

We learn two things from the regression output

Regression coefficient reports effect of x on probability to be placed one category higher

- Effect in logodds: `0.084`
- We can backtransform to one unit increase in x: $(exp(\beta) - 1) \times 100 =$ 9% increase in likelihood of a higher category.

$\Rightarrow$ *Hypothesis testing as in a binomial logit*

We learn two things from the regression output

We have one intercept per cutpoint

- e.g.: intercept of passing from 1 to 2 is `-2.012`
- e.g.: intercept is reported as significant (with standard errors)

⇒ *The model does a fair job in distinguishing between categories.*

Predicted scenarios

We interpret predicted effect by choosing one level of x and one category (two cutpoints) of y: What is the probability of m?

$$Pr(y_i = m) = \frac{exp(\tau_m - \beta x_i)}{1 + exp(\tau_m - \beta x_i)} - \frac{exp(\tau_{m-1} - \beta x_i)}{1 + exp(\tau_{m-1} - \beta x_i)} \qquad (2)$$

Example

Let's choose low-income respondents (x = 1) and category 3 (diff between cutpoints 2 and 3)

```
z = mod.ord$zeta
x = 1

logodds1 <- z[3] - coefficients(mod.ord) * x
logodds2 <- z[3-1] - coefficients(mod.ord) * x
## Probabilities
p1 <- exp(logodds1)/(1 + exp(logodds1)) #3/4 or lower
p2 <- exp(logodds2)/(1 + exp(logodds2)) #2/3 or lower
## Difference between cutpoints
p1 - p2 #cat 3
```

An example

```
paste(round((p1)*100),
"% of low-income respondents are predicted to answer x = 3 ('r
```

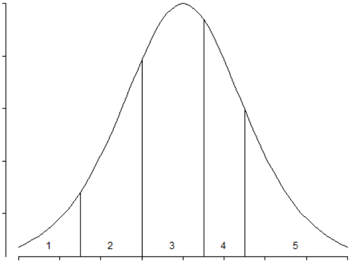
[1] "80 % of low-income respondents are predicted to answer x = 3 ('neutral') or lower to the question of whether they support redistribution."

```
paste(round((p1-p2)*100),
"% of low-income respondents are predicted to answer x = 3 ('r
```

[1] "25 % of low-income respondents are predicted to answer x = 3 ('neutral')."

Two ways of viewing the slicing

We can report the probability (e.g. `0.25`) of ending up between two cutpoints, or the *cumulative* probability (e.g. `0.8`) to be below each point.



Parallel regressions approach: for assessment

Parallel regressions approach

The parallel regression approach is useful to understand how the model is estimated

- The y is recoded into $m-1$ dummy variables indicating if $y \leq m$
- Run a series of regressions where all β are fixed (the same).

An example of parallel regressions

Recode into dummies

The dummies flag cases below a cumulative threshold of *outcomes*

```
##
df$ut1 <- ifelse(df$Utjevn > 1, 1 , 0) #2 or above
df$ut2 <- ifelse(df$Utjevn > 2, 1 , 0) #3 or above
df$ut3 <- ifelse(df$Utjevn > 3, 1 , 0) #4 or above
df$ut4 <- ifelse(df$Utjevn > 4, 1 , 0) #5
```

$\Rightarrow$ *The model then runs 4 regressions where β reports an aggregated value from all 4 coefficients (think: weigted mean).*

How good is our model?

The basic assumption

The basic assumption is that all parallel regressions have (about) the same regression coefficient

- Check the mean of the predictor for each value of y. Does it trend?

```
##        1        2        3        4        5
## 4.742857 5.547059 5.438017 6.205607 6.571429
```

- Run parallel regressions without contstraint on β. Are they similar?

Example: testing the parallell regressions assumtiopn

Run four regressions

Let's examplify with the parallel regressions without fixed β:

```
##Parallel regressions:
mod1 <- glm(ut1 ~ Inntekt, df, family = "binomial")
mod2 <- glm(ut2 ~ Inntekt, df, family = "binomial")
mod3 <- glm(ut3 ~ Inntekt, df, family = "binomial")
mod4 <- glm(ut4 ~ Inntekt, df, family = "binomial")
```

Compare coefficients from four regressions

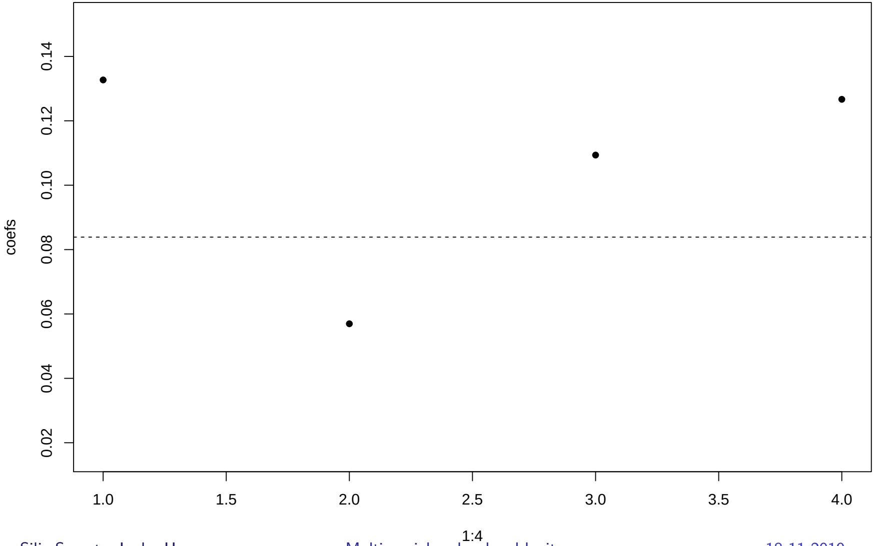
```
##
## ===================================================
##                          Dependent variable:
##                     -------------------------------
##                       ut1     ut2      ut3      ut4
##                       (1)     (2)      (3)      (4)
## ---------------------------------------------------
## Inntekt             0.133**  0.057*  0.109***  0.127
##                     (0.067)  (0.035) (0.039)  (0.101)
##
## Constant            1.772*** -0.155  -1.629*** -4.204***
##                     (0.367)  (0.216) (0.259)  (0.711)
##
## ---------------------------------------------------
## Observations          447     447      447      447
## Log Likelihood      -120.685 -306.937 -257.053 -61.452
## Akaike Inf. Crit.   245.370  617.875  518.105  126.903
## ===================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

Coefficient should be a weighted average from four regressions

These βs are weighted by the number of observations in each category:

```
##
##    1   2   3   4   5
##   40 174 127 108  14
```

We can plot the βs for comparison:



When is it smart to run an ordered logit?

- You have few categories
- Fairly equal spread of observations between categories

Discrete choice models

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

Multinomial logistic regression

Two conceptions of multinomial regression

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

Main assumption: IIA

Independence of irrelevant alternatives:

- there are no choices beyond what is modelled
- 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 vs. unrestricted model
- There should be no difference (X^2-test)

Prediction testing

- confusion matrix (Proportiono of correct predictions: $\frac{sumofdiagonal}{nobservations}$)
- one-versus-all
- ROC curve and separation plots

$\Rightarrow$ *as in binomial regression*

Interpretation

All the possibilities of the binomial logit are open:

- The regression table
- Predicted probabilities (and comparisons/scenarios)

Specific visual interpretations

- The three dimensional simplex (if $M = 3$)
- The ternary plot: a sort of scatterplot for predicted probabilities $\Rightarrow$ *Illustrates tradeoffs*

The conditional logit

From the chooser's perspectives

The conditional logit holds the chooser constant, and consider alternative choices

- Parallel regressions approach: a logit in a choice set
- One set of parameters, no intercept
- Long data format (observation = choice in individual)

Mixing choosers and choices

The mixed conditional logit makes an interaction effect between choice-set variables and choice variables.