Ordinal Logistic Regression

POST 8000 – Foundations of Social Science Research for Public Policy

Steven V. Miller

Department of Political Science

Goal for Today

Discuss ordinal logistic regression (i.e. what to do when your DV is ordered-categorical).

Order Without Precision

You may want to explain ordered-categorical responses like:

- Patient quality of life (excellent, good, fair, poor)
- Political philosophy (very liberal:very conservative on a five-point scale)
- Government spending (too low, about right, too high)
- Likert items (strongly agree:strongly disagree, five point-scale)

Problem: order is implied, but without precision.

- A vague magnifier of "very" separates "very liberal" from "liberal."
- "Strongly" separates "strongly agree" from "agree."

Scales are discrete, which will again break the assumptions of OLS.

Simulated Ordinal Data

Observe again this simulated ordinal data, D.

```
D %>% select(y, x1, x2)
```

```
## # A tibble: 10,000 x 3
##          y      x1     x2
##      <dbl>   <dbl>  <dbl>
##  1       2 -0.997   0.689
##  2       5  0.722   1.08
##  3       1 -0.617  -0.221
##  4       5  2.03    0.485
##  5       5  1.07    0.196
##  6       5  0.987   0.285
##  7       1  0.0275 -0.914
##  8       4  0.673  -1.07
##  9       4  0.572  -0.395
## 10       5  0.904  -1.86
## # ... with 9,990 more rows
```

Simulated Ordinal Data

This simple `D` data frame is simulated where:

- y is a five-item ordered-categorical variable (similar to a Likert item).
- $x1$ and $x2$ are random-normal with means 0 and SDs of 1.
- The effect of $x1$ on y is 1.
- The effect of $x2$ on y is .5.

Importantly: y is sampled from individual probabilities for each observation.

- Probability is determined by cumulative logits.
- tl;dr: data aren't simulated as an OLS model.

Simulated Ordinal Data

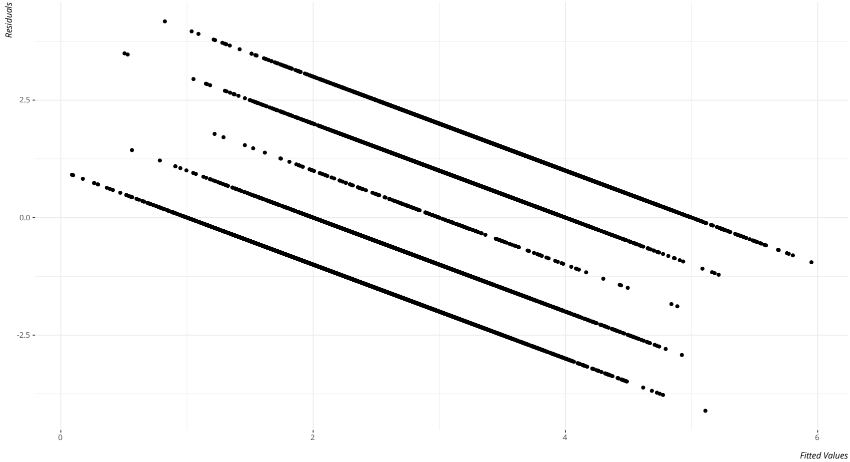
```r
M1 <-lm(y ~ x1 + x2, D)
broom::tidy(M1) %>%
  mutate_if(is.numeric,~round(.,2))%>%
  kable(.,"markdown")
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.98	0.01	204.11	0
x1	0.75	0.01	51.91	0
x2	0.36	0.01	24.32	0

Not bad, but not correct.

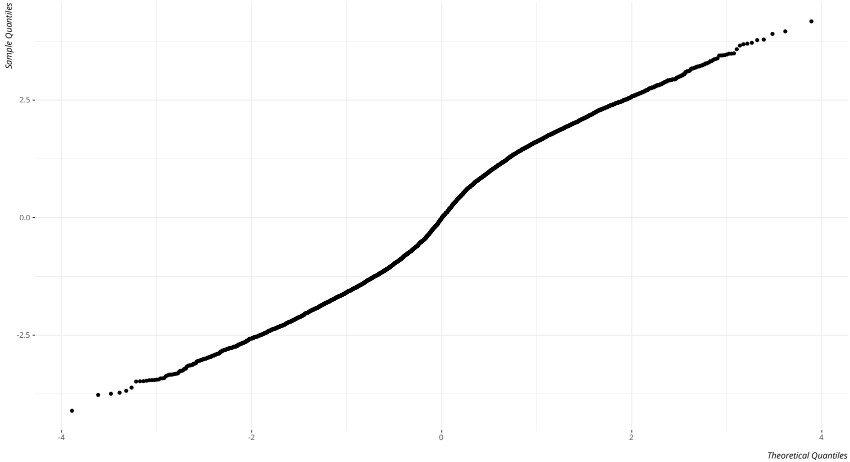
The Fitted-Residual Plot from the OLS Model We Just Ran

Patterns shouldn't emerge from a fitted-residual plot, but they will with discrete DVs like this.



The Q-Q Plot from the OLS Model We Just Ran

The Q-Q plot is a little more sanguine about what we did, but notice that bend in the middle.



The Right Tool for the Right Job

```r
require(ordinal)
D$y_ord = ordered(D$y)
M2 <- clm(y_ord ~ x1 + x2, data = D)
broom::tidy(M2) %>%
    mutate_if(is.numeric, ~round(., 2)) %>%
  filter(coefficient_type == "beta") %>%
    kable(., "markdown")
```

term	estimate	std.error	statistic	p.value	coefficient_type
x1	1.01	0.02	46.16	0	beta
x2	0.47	0.02	24.02	0	beta

What We Just Did

This was an ordinal logistic regression. Related names/terms:

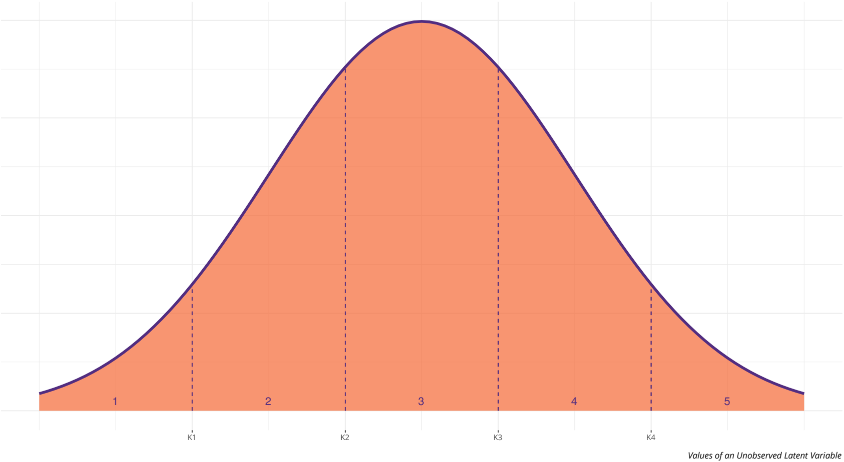
- “Proportional odds logistic regression”
- “Cumulative link model”

Conceptualizing an Ordinal Logistic Regression

1. There is an observed ordinal variable Y.
2. There is an underlying, *un*observed latent variable Y^* that is continuous.

An Ordered Response and its Latent Variable

The idea is there is an underlying latent variable, but we're only observing five collapsed values of it.



Conceptualizing an Ordinal Logistic Regression

Y^* has various threshold points κ.

- The value of Y is observed whether you've passed a certain threshold.

For example, when number of categories = 5:

- Y_i = 1 if Y_i^* is $\leq \kappa_1$
- Y_i = 2 if $\kappa_1 \leq Y_i^* \leq \kappa_2$
- Y_i = 3 if $\kappa_2 \leq Y_i^* \leq \kappa_3$
- Y_i = 4 if $\kappa_3 \leq Y_i^* \leq \kappa_4$
- Y_i = 5 if Y_i^* is $\geq \kappa_4$

Conceptualizing an Ordinal Logistic Regression

In the population, the continuous latent variable Y^* is equal to

$$Y^* = \sum_{k=1}^{K} \beta_k X_{ki} + \epsilon_i = Z_i + \epsilon_i$$

Of note: that random disturbance term has a standard logistic distribution.

Conceptualizing an Ordinal Logistic Regression

The ordinal logistic regression estimates part of the above.

$$Z_i = \sum_{k=1}^{K} \beta_k X_{ki} = E(Y_i^*)$$

Because Z is not a perfect measure of Y^*, there are prediction errors.

- But, knowing the distribution of the error term, you can calculate those probabilities.

Conceptualizing an Ordinal Logistic Regression

What you need to estimate from this design: the κs, the βs in order to compute $Z_i = \sum_{k=1}^{K} \beta_k X_{ki}$.

- Note: there is no intercept term in this extension of the logistic model.

For shorthand we teach this as a “parallel lines” model.

- The ordinal logistic regression estimates one equation over all levels of the response variable.
- The estimate that emerges communicates the natural logged odds of going up (or down) a “level”, all else equal.

Assumptions/Diagnostics

Ordinal logistic regression is a GLM, so much still applies from logistic regression.

- e.g. deviance, log-likelihood, MLE

One important assumptions: parallel lines (proportional odds).

- The slope estimate between each pair of outcomes across two response levels are assumed to be the same regardless of which partition we consider.

Assumptions/Diagnostics

There are no shortage of tests for this:

- Brant test
- LR test
- Wald test

Implication: violating the assumption means the ordinal logistic model is too restrictive.

- Consider a multinomial GLM instead.

Conclusion

If your DV is ordered-categorical, considered an ordinal logistic regression.

- OLS works much better with ordered-categorical DVs than binary DVs, but that's less the point.
  - Feel free to compare OLS with it, though OLS would be "wrong" in a non-trivial way.
- Think of these as situations where Y^* is assumed but only a finite Y iis observed.

Ordinal logistic regression better models the nature of the data, but:

- check for parallel lines/proportional odds.
- coefficient interpretation is a little weirder.
  - my take: be prepared to communicate quantities of interest as probabilities.

Table of Contents