Proportional odds and Probit regression

Rebecca C. Steorts (slide adaption from Maria Tacket) and material Chapters 6 and 7 of McNulty (2021).

Computing set up

```
library(tidyverse)
library(tidymodels)
library(knitr)
library(margins)

knitr::opts_chunk$set(fig.width = 8,
                      fig.asp = 0.618,
                      fig.retina = 3,
                      dpt = 300,
                      out.width = "70%",
                      fig.align = "center")

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Anouncements

1. Last homework due on Wednesday, April 16th at 5 PM.
2. Quiz 4 (last quiz) due on LDOC at 5 PM.
3. Rest of semester will finish on working on quiz 4.

Lecture today

1. Quickly go through rest of class material so you can finalize last homework.
2. Go over quiz 4. Spend class time Wednesday and Monday working on this.
3. Use LDOC to finalize this and polish if needed or ask questions.
4. Other things?

Learning goals

- Introduce proportional odds and probit regression models
- Understand how these models are related to logistic regression models
- Interpret coefficients in context of the data
- See how these models are applied in research contexts

Notes based on Chapters 6 and 7 of McNulty (2021) unless stated otherwise.

Proportional odds models

Predicting ED wait and treatment times

Ataman and Sarıyer (2021) use ordinal logistic regression to predict patient wait and treatment times in an emergency department (ED). The goal is to identify relevant factors that can be used to inform recommendations for reducing wait and treatment times, thus improving the quality of care in the ED.

Data: Daily records for ED arrivals in August 2018 at a public hospital in Izmir, Turkey.
Article: https://www.sciencedirect.com/science/article/abs/pii/S0735675721001698?via%3Dihub

Predicting ED wait and treatment times

Response variables:

- `Wait time`:
  - Patients who wait less than 10 minutes
  - Patients whose waiting time is in the range of 10 - 60 minutes
  - Patients who wait more than 60 minutes
- `Treatment time`:
  - Patients who are treated for up to 10 minutes
  - Patients whose treatment time is in the range of 10 - 120 minutes
  - Patients who are treated for longer than 120 minutes

Predicting ED wait and treatment times

Predictor variables:

- `Gender`:
  - Male
  - Female
- `Age`:
  - 0 - 14
  - 15 - 64
  - 65 - 84
  - $\geq$ 85
- `Arrival mode`:
  - Walk-in
  - Ambulance
- `Triage level`:
  - Red (urgent)
  - Green (non-urgent)
- `ICD-10 diagnosis`: Codes specifying patient's diagnosis

Ordered vs. unordered variables

Categorical variables with 3+ levels

Unordered (Nominal)

- Voting choice in election with multiple candidates
- Type of cell phone owned by adults in the U.S.
- Favorite social media platform among undergraduate students

Ordered (Ordinal)

- Wait and treatment times in the emergency department
- Likert scale ratings on a survey
- Employee job performance ratings

Proportional odds model

Let Y be an ordinal response variable that takes levels $1, 2, \ldots, J$ with associated probabilities $p_1, p_2, \ldots, p_J$

Proportional odds model

Let Y be an ordinal response variable that takes levels $1, 2, \ldots, J$ with associated probabilities $p_1, p_2, \ldots, p_J$

The **proportional odds model** can be written as the following:

$$\log\left(\frac{P(Y \leq 1)}{P(Y > 1)}\right) = \beta_{01} - \beta_1 x_1 - \cdots - \beta_p x_p$$

$$\log\left(\frac{P(Y \leq 2)}{P(Y > 2)}\right) = \beta_{02} - \beta_1 x_1 - \cdots - \beta_p x_p$$

$$\ldots$$

$$\log\left(\frac{P(Y \leq J-1)}{P(Y > J-1)}\right) = \beta_{0J-1} - \beta_1 x_1 - \cdots - \beta_p x_p$$

What does β_{01} mean? What does β_1 mean?

Proportional odds model

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \beta_{0k} - \beta_1 x_1 - \cdots - \beta_p x_p$$

Suppose $\beta_1 > 0$.

- Then as x_1 increases, the $\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right)$ decreases since we are subtracting β_1.
- This means that the odds of being in a lower category decrease.
- Thus, the odds of being in a higher category increase.
- To summarize, $\beta_1 > 0$ is associated with increased **log-odds** of being in a **higher** category of Y

Proportional odds model

Let's consider one portion of the model:

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \beta_{0k} - \beta_1 x_1 - \cdots - \beta_p x_p$$

Proportional odds model

Let's consider one portion of the model:

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \beta_{0k} - \beta_1 x_1 - \cdots - \beta_p x_p$$

- The response variable is $logit(Y \leq k)$, the log-odds of observing an outcome less than or equal to category k.

Proportional odds model

Let's consider one portion of the model:

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \beta_{0k} - \beta_1 x_1 - \cdots - \beta_p x_p$$

- The response variable is $logit(Y \leq k)$, the log-odds of observing an outcome less than or equal to category k.
- $\beta_j > 0$ is associated with increased **log-odds** of being in a **higher** category of Y

Proportional odds model

Let's consider one portion of the model:

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \beta_{0k} - \beta_1 x_1 - \cdots - \beta_p x_p$$

- The response variable is $logit(Y \leq k)$, the log-odds of observing an outcome less than or equal to category k.
- $\beta_j > 0$ is associated with increased **log-odds** of being in a **higher** category of Y
  - e^{β_j} associated with an increased **odds** of being in a **higher** category of Y

Proportional odds model

Let's consider one portion of the model:

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \beta_{0k} - \beta_1 x_1 - \cdots - \beta_p x_p$$

- The response variable is $logit(Y \leq k)$, the log-odds of observing an outcome less than or equal to category k.
- $\beta_j > 0$ is associated with increased **log-odds** of being in a **higher** category of Y
  - e^{β_j} associated with an increased **odds** of being in a **higher** category of Y
- Effect of one unit increase in x_j is the same regardless of which category of Y

Example

Suppose you have an ordinal outcome variable, `Satisfaction''`, `with categories:`Low'', `Medium''`,High'', and a predictor "Income''. Consider

$$\log\left(\frac{P(Y \leq \text{Medium})}{P(Y > \text{Medium})}\right) = -0.5 + 0.2 \times \text{Income}$$

- The coefficient for **Income** is 0.2. This means that for each one-unit increase in Income, the log-odds of being in a higher satisfaction category increase by 0.2.
- In terms of odds, $e^{0.2} \approx 1.22$, which means that for each additional unit increase in Income, the odds of being in a higher category of Satisfaction (High versus Low or Medium) increase by a factor of 1.22.

Connection with R

- The `polr()` function in R is used for fitting ordered logistic regression models.
- Regression coefficients in the polr model represent the change in the log-odds of being in a higher category of the outcome variable for a one-unit increase in the predictor variable, holding other predictors constant.
- In summary, the R function fits the model given above, which is why the coefficient interpretations may appear flipped. It is essential to check functions and your model formulation match!

Emergency Department Study

Paper: https://www.sciencedirect.com/science/article/pii/S0735675721001698?via=ihub

- Retrospective data on 37,711 patients arriving at the ED of a large urban hospital were examined.
- Ordinal logistic regression models were proposed to identify factors causing increased waiting and treatment times and classify patients with longer waiting and treatment times.
- In this application, the model was fit using the one assumed in the slides.

Effect of arrival mode on waiting time

M.G. Ataman and G. Sarıyer

Table 5
OLR models results

Input variable	OLR model for waiting time			
	Parameter estimate	*p*-value	95% confidence interval	
			Lower bound	Upper bound
Gender	−0.022	0.261	−0.061	0.016
Age	−0.116	0.000	−0.154	−0.079
Arrival mode	−3.398	0.000	−3.616	−3.180
Triage level	0.016	0.153	−0.006	0.037
ICD-10 diagnosis	−0.067	0.000	−0.071	−0.063
	Model fitting information: Chi-square = 3740.277; *p*-value: 0.000			
	Model summary: Cox & Snell R square = 0.194; Nagelkerke R square = 0.207			

Figure 1: Waiting time model output from Ataman and Sarıyer (2021)

Question

The variable `arrival mode` has two possible values: ambulance and walk-in. Describe the effect of arrival mode (covariate) on waiting time (three levels). Note: The baseline category is for `arrival mode` is walk-in.

Hint: Recall how the model is written and our interpretation of the coefficients.

Solution

- The p-value (0.0000) is small indicating `arrival mode` is a statistically significant predictor of wait time after adjusting for the other factors.
- The coefficient for `arrival mode = ambulance` (compared to walk-in) is -3.398. In terms of odds, $e^{-3.398} \approx 0.033$.
- If the patient arrived via an ambulance, the odds of being in a higher wait time category **decrease** by a factor of 0.033, holding the other factors constant.
- Thus, arriving by ambulance **decreases** the likelihood of being classified in the longer waiting groups, after adjusting for all other factors.

Effect of arrival mode on waiting and treatment time

Consider the full output with the ordinal logistic models for wait and treatment times.

Table 5
OLR models results

Input variable	OLR model for waiting time				OLR model for treatment time			
	Parameter estimate	*p*-value	95% confidence interval		Parameter estimate	*p*-value	95% confidence interval	
			Lower bound	Upper bound			Lower bound	Upper bound
Gender	−0.022	0.261	−0.061	0.016	0.041	0.056	−0.001	0.084
Age	−0.116	0.000	−0.154	−0.079	0.151	0.000	0.111	0.190
Arrival mode	−3.398	0.000	−3.616	−3.180	1.215	0.000	1.095	1.335
Triage level	0.016	0.153	−0.006	0.037	−0.950	0.000	−0.973	−0.926
ICD-10 diagnosis	−0.067	0.000	−0.071	−0.063	0.054	0.000	0.049	0.058
	Model fitting information: Chi-square = 3740.277; *p*-value: 0.000 Model summary: Cox & Snell R square = 0.194; Nagelkerke R square = 0.207				Model fitting information: Chi-square = 10,504.755; *p*-value: 0.000 Model summary: Cox & Snell R square = 0.343; Nagelkerke R square = 0.382			

Figure 2: Waiting and treatment time model output from Ataman and Sarıyer (2021).

Question

Now, consider the effect of `arrival mode` on waiting time and treatment time.

Solution

- The p-value (0.0000) is small indicating `arrival mode` is a statistically significant predictor of treatment time after adjusting for the other factors.
- The coefficient for `arrival mode = ambulance` (compared to walk-in) is 1.125. In terms of odds, $e^{1.125} \approx 3.08$.
- If the patient arrived via an ambulance, the odds of being in a higher treatment category **increases** by a factor of 3.08, holding the other factors constant.
- Thus, arriving by ambulance **increases** the likelihood of being classified in the longer treatment time group (after adjusting for all other factors).

Summary

Arriving by ambulance decreases the likelihood of being classified in the longer waiting time groups, but increases the likelihood of being classified in the longer treatment time groups.

Fitting proportional odds models in R

Fit proportional odds models using the `polr` function in the **MASS** package:

```
proportional_model <-
  polr(Y ~ x1 + x2 + x3, data = my_data)
```

Multinomial logistic model

Suppose the outcome variable Y is categorical and can take values $1, 2, \ldots, K$ such that

$$P(Y = 1) = p_1, \ldots, P(Y = K) = p_K \qquad \text{and} \qquad \sum_{k=1}^{K} p_k = 1$$

Multinomial logistic model

Suppose the outcome variable Y is categorical and can take values $1, 2, \ldots, K$ such that

$$P(Y = 1) = p_1, \ldots, P(Y = K) = p_K \qquad \text{and} \qquad \sum_{k=1}^{K} p_k = 1$$

Choose baseline category. Let's choose $Y = 1$. Then

Multinomial logistic model

Suppose the outcome variable Y is categorical and can take values $1, 2, \ldots, K$ such that

$$P(Y = 1) = p_1, \ldots, P(Y = K) = p_K \qquad \text{and} \qquad \sum_{k=1}^{K} p_k = 1$$

Choose baseline category. Let's choose $Y = 1$. Then

$$\log\left(\frac{P(Y = 2)}{P(Y = 1)}\right) = \beta_{02} - \beta_{12}x_1 - \cdots - \beta_{p2}x_p$$

$$\log\left(\frac{P(Y = 3)}{P(Y = 1)}\right) = \beta_{03} - \beta_{13}x_1 - \cdots - \beta_{p3}x_p$$

$$\ldots$$

$$\log\left(\frac{P(Y = K)}{P(Y = 1)}\right) = \beta_{0K} - \beta_{1K}x_1 - \cdots - \beta_{pK}x_p$$

Interpretation and Example

1. The model estimates the log-odds of each category relative to a baseline category, allowing for the prediction of probabilities for all categories.

2. The first equation looks to see how a one unit change in each coefficient changes the log-odds of going from Category 2 to Category 1 (baseline). We can look at the interpretation as the log-odds or the odds (just as we did for logistic regression).

3. Because every coefficient in this model refers back to the baseline category, you need to think carefully through which reference category to choose. You have to think what set of comparison is the most relevant to the application.

4. See https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/ regarding an example for multinomial regression (and the interpretation of the coefficients).

5. Another resource is available here: https://bookdown.org/sarahwerth2024/CategoricalBook/probit-regression-r.html

Multinomial logistic vs. proportional odds

How is the proportional odds model similar to the multinomial logistic model? How is it different? What is an advantage of each model? What is a disadvantage?

Solution

Solution

Probit regression

Impact of nature documentary on recycling

Ibanez and Roussel (2022) conducted an experiment to understand the impact of watching a nature documentary on pro-environmental behavior. The researchers randomly assigned the 113 participants to watch an video about architecture in NYC (control) or a video about Yellowstone National Park (treatment). As part of the experiment, participants were asked to dispose of their headphone coverings in a recycle bin available at the end of the experiment.

Article: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0275806

Impact of nature documentary on recycling

Response variable: Recycle headphone coverings vs. not

Predictor variables:

- Age
- Gender
- Student
- Made donation to environmental organization in previous part of experiment
- Environmental beliefs measured by the new ecological paradigm scale (NEP)

Probit regression

Let Y be a binary response variable that takes values 0 or 1, and let $p = P(Y = 1|x_1, \ldots, x_p)$

$$probit(p) = \Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where Φ^{-1} is the inverse normal distribution function.

Probit regression

Let Y be a binary response variable that takes values 0 or 1, and let $p = P(Y = 1|x_1, \ldots, x_p)$

$$probit(p) = \Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where Φ^{-1} is the inverse normal distribution function.

The outcome is the z-score at which the cumulative probability is equal to p

- e.g. $probit(0.975) = \Phi^{-1}(0.975) = 1.96$

Interpretation

- $\hat{\beta}_j$ is the estimated change in z-score for each unit increase in x_j, holding all other factors constant.
- This is a fairly clunky interpretation, so the **(average) marginal effect** of x_j is often interpreted instead
- The marginal effect of x_j is essentially the change the probability from variable x_j

Impact of nature documentary

VARIABLES	Probit 0/1 *Likelihood*	Marginal effects *Probability points*
Nature (T2)	0.841***	0.279**
	(0.318)	(0.095)
Urban (T1)	*Ref.*	
Donation (Yes)	-0.041	-0.013
	(0.323)	(0.107)
Gender (Male)	0.064	0.021
	(0.271)	(0.090)
Age	-0.083**	-0.028**
	(0.036)	(0.011)
Student	-0.199	-0.066
	(0.485)	(0.161)
NEP-High	1.500***	0.478***
	(0.402)	(0.091)
*Nature (T2) * NEP-High*	-1.016*	
	(0.576)	
Constant	1.389	
	(1.104)	
LL	-66.157	
LR Chi^2(7)	23.62***	
Pseudo R^2	0.152	
Number of observations	113	
Session controls	Yes	

Standard errors in parentheses; significant levels

*** p<0.01

** p<0.05

* p<0.1.

https://doi.org/10.1371/journal.pone.0275806.t006

Figure 3: Recycling model from Ibanez and Roussel (2022)

Interpret the effect of watching the nature documentary `Nature (T2)` on recycling. Assume NEP is low, `NEP-High` $= 0$.

Solution

Interpret the effect of watching the nature documentary `Nature (T2)` on recycling. Assume NEP is low, `NEP-High` $= 0$.

- Participants exposed to the natural setting (Nature (T2), 0.841***) are more likely to recycle than those exposed to the urban setting (T1).
- This is reflected in the marginal effects in terms of percentage points: the probability of recycling rises under exposure to nature (Nature (T2), 0.279***) compared with the urban exposure treatment.

(See page 13, Ibanez and Roussel (2022).)

Probit vs. logistic regression

Pros of probit regression:

- Some statisticians like assuming the normal distribution over the logistic distribution.
- Easier to work with in more advanced settings, such as multivariate and Bayesian modeling

Probit vs. logistic regression

Pros of probit regression:

- Some statisticians like assuming the normal distribution over the logistic distribution.
- Easier to work with in more advanced settings, such as multivariate and Bayesian modeling

Cons of probit regression:

- Z-scores are not as straightforward to interpret as the outcomes of a logistic model.
- We can't use odds ratios to describe findings.
- It's more mathematically complicated than logistic regression.
- It does not work well for response variable with 3+ categories

List adapted from Categorical Regression.

Fitting probit regression models in R

Fit probit regression models using the `glm` function with `family = binomial(link = probit)`.

Calculate marginal effects using the `margins` function from the **margins** R package.

```
margins(my_model, variables = "my_variables")
```

Wrap up GLM for independent observations

Wrap up

- Covered fitting, interpreting, and drawing conclusions from GLMs
  - Looked at Poisson, Negative Binomial, and Logistic, Proportional odds, and Probit models in detail
- Used Pearson and deviance residuals to assess model fit and determine if new variables should be added to the model
- Addressed issues of overdispersion and zero-inflation
- Used the properties of the exponential family to identify canonical link function for any GLM and additional properties via the log-partition function.

Wrap up

- Covered fitting, interpreting, and drawing conclusions from GLMs
  - Looked at Poisson, Negative Binomial, and Logistic, Proportional odds, and Probit models in detail
- Used Pearson and deviance residuals to assess model fit and determine if new variables should be added to the model
- Addressed issues of overdispersion and zero-inflation
- Used the properties of the exponential family to identify canonical link function for any GLM and additional properties via the log-partition function.

Everything we've done thus far as been under the assumption that the observations are *independent*. Looking ahead we will consider models for data with **dependent (correlated) observations**.

References

Ataman, Mustafa Gökalp, and Görkem Sarıyer. 2021. "Predicting Waiting and Treatment Times in Emergency Departments Using Ordinal Logistic Regression Models." *The American Journal of Emergency Medicine* 46: 45–50.

Ibanez, Lisette, and Sébastien Roussel. 2022. "The Impact of Nature Video Exposure on Pro-Environmental Behavior: An Experimental Investigation." *Plos One* 17 (11): e0275806.

McNulty, Keith. 2021. *Handbook of Regression Modeling in People Analytics: With Examples in r and Python*. CRC Press.