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")
```

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.
Click here to access the article on Canvas.

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?

Solution: What does β_{01} mean?

The term β_{01} is the **intercept** in the first cumulative log-odds equation:

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

This represents the **log-odds** of the outcome $Y \leq 1$ (i.e., the probability that Y is either 1 or less) compared to $Y > 1$ (i.e., the probability that Y is greater than 1), when all covariates $x_1, x_2, \ldots, x_p$ are equal to zero.

In simpler terms, β_{01} reflects the **baseline log-odds** of being in the lower category of Y (i.e., $Y \leq 1$) versus being in a higher category, when no explanatory variables $(x_1, x_2, \ldots, x_p)$ are present. It essentially sets the starting point for the relationship between the covariates and the odds of being in a lower ordinal category.

Solution: What does β_1 mean?

The term β_1 is the coefficient for the first explanatory variable x_1 in all of the log-odds equations:

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \beta_{0j} - \beta_1 x_1 - \cdots - \beta_p x_p \quad \text{for each } j = 1, 2, \ldots, J-1.$$

The coefficient β_1 represents the **change in the log-odds** of being in a lower or equal category (i.e., $Y \leq j$) versus being in a higher category (i.e., $Y > j$) for a **one-unit increase** in x_1, holding all other covariates constant.

- A **positive** β_1 means that as x_1 increases, the odds of being in a lower category (or a category less than or equal to j) increase, suggesting that higher values of x_1 are associated with being in higher ordinal categories.
- A **negative** β_1 means that as x_1 increases, the odds of being in a lower category decrease, suggesting that higher values of x_1 are associated with being in higher ordinal categories.

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 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". The proportional odds model might give the following coefficient:

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

- 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 (compared to being in a lower 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 (Medium or High vs. Low) increase by a factor of 1.22.

Effect of arrival mode on waiting time

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)

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

Solution

- Arrival mode has two possible values: “ambulance” and “walk-in”.
- Baseline category is “walk-in”. This means that the model’s intercept is associated with “walk-in”, and the coefficient for “ambulance” describes how the log-odds change relative to the “walk-in” category.
- The coefficient for “arrival mode” (for ambulance) is -3.398.
- For individuals arriving by ambulance (compared to those who walk in), the log odds of waiting time decrease by 3.398.
- A negative coefficient suggests that arriving by ambulance is associated with a lower log-odds of waiting time compared to walking in.

Effect of triage level

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).

Triage levels have three possible values: “trauma”, “red,” and “yellow.” Use the results from both models to describe the effect of triage level on waiting and treatment times. Note: The baseline category is green.

Solution

- Triage Levels: “trauma”, “red”, and “yellow” (with green as the baseline category).
- The baseline category is green, so the interpretation of the coefficient applies to the comparison between the triage levels “trauma”, “red”, and “yellow” relative to “green”.
- The coefficient for triage level is 0.016 (for waiting time); the coefficient for triage level is -0.950 for treatment time.
- The coefficient of 0.016 implies that patients with higher triage levels (compared to the “green” baseline) have a slightly higher likelihood of experiencing longer waiting time.
- The coefficient of -0.950 implies that patients with higher triage levels (compared to the “green” baseline) have a lower likelihood of experiencing a longer treatment time.

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

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

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

$$\ldots$$

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

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

How is the proportional odds model similar to the multinomial logistic model?

1. Both models are used when the dependent variable is categorical with more than two categories.
2. Both models aim to understand how independent variables (predictors) influence the probability of observing one of the different categories of the dependent variable.
3. Both models are appropriate for analyzing multicategory responses (e.g., categories like “low”, “medium”, “high”).
4. Both models assume a log-odds framework. The relationship between the independent variables and the dependent variable is expressed in terms of log-odds. In this way, both models allow for non-linear relationships between the predictors and the response variable.

Solution

How is the proportional odds model different from the multinomial logistic model?

1. In the proportional odds model, the dependent variable is ordinal, meaning it has categories with a natural order or ranking.
2. In the multinomial logistic model, the dependent variable is nominal, which does not assume any ordering of the variables.

Solution

What is a disadvantage of multinomial logistic model compared to the proportional odds model?

The multinomial logistic model requires more parameters to be estimated than the proportional odds model, particularly when the dependent variable has more than two categories.

- The multinomial logistic models separate sets of coefficients for each category compared to a baseline category. For an outcome for K categories, the multinomial logistic model estimates $K - 1$ coefficients.
- The proportional odds model only estimates one set of coefficients, which is more parsimonious and computationally more efficient.

start here

start here and possibly break these up into multiple lectures.

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.

Click here to access the article on Canvas.

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 performing a green deed 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")
```

Ideology vs. issue statements

Let's look at the model using ideology and party ID to explain the number of issue statements by politicians.

We will use probit regression for the "hurdle" part of the model - the likelihood a candidate comments on at least one issue (`has_issue_stmt`)

Ideology vs. issue statements

Let's look at the model using ideology and party ID to explain the number of issue statements by politicians.

We will use probit regression for the "hurdle" part of the model - the likelihood a candidate comments on at least one issue (`has_issue_stmt`)

```
hurdle_probit <- glm(has_issue_stmt ~
                       ideology + democrat,
                data = politics,
                family = binomial(link = probit))
```

See Section 4.11.2 of Roback and Legler (2021) for more detail about the data.

Hurdle (using probit regression)

```
tidy(hurdle_probit) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.272	0.117	10.829	0.000
ideology	0.262	0.089	2.926	0.003
democrat1	0.149	0.180	0.827	0.408

```
margins(hurdle_probit)
```

```
## Average marginal effects

## glm(formula = has_issue_stmt ~ ideology + democrat, fami

##  ideology democrat1
##   0.04071   0.02333
```

Interpret the effect of `democrat` on commenting on at least one issue.

Hurdle (using logistic regression)

```
hurdle_logistic <- glm(has_issue_stmt ~ ideology + democrat
                       data = politics,
                       family = binomial)

tidy(hurdle_logistic) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.127	0.225	9.465	0.000
ideology	0.575	0.167	3.446	0.001
democrat1	0.428	0.349	1.225	0.221

Probit vs. logistic models

Probit model

term	estimate
(Intercept)	1.272
ideology	0.262
democrat1	0.149

Logistic model

term	estimate
(Intercept)	2.127
ideology	0.575
democrat1	0.428

Suppose there is democratic representative with ideology score -2.5. Based on the probit model, what is the probability they will comment on at least one issue? What is the probability based on the logistic model?

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 one-parameter exponential family to identify the best link function for any GLM

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 one-parameter exponential family to identify the best link function for any GLM

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.

Roback, Paul, and Julie Legler. 2021. *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. CRC Press.