Lecture 3 - Generalized Linear Models: Ordinal Logistic Regression

Contents

- Today's Learning Goals
- Loading Libraries
- 1. The Need for an Ordinal Model
- 2. College Juniors Dataset
- 3. Exploratory Data Analysis
- 4. Data Modelling Framework
- 5. Estimation
- 6. Inference
- 7. Coefficient Interpretation
- 8. Predictions
- 9. Model Selection
- 10. Non-proportional Odds
- 11. Wrapping Up on Categorical Regression Models

Today's Learning Goals

By the end of this lecture, you should be able to:

- Outline the modelling framework of the Ordinal Logistic regression.
- Explain the concept of proportional odds.
- Fit and interpret Ordinal Logistic regression.

- Use the Ordinal Logistic regression for prediction.
- Explain the concept of non-proportional odds.
- Assess the assumption of proportional odds via the Brant-Wald test.
- Contrast the Ordinal Logistic regression models under proportional and non-proportional odds assumptions.

Loading Libraries

```
options(repr.matrix.max.rows = 6)
source("../scripts/support_functions.R")
library(VGAM)
library(cowplot)
library(tidyverse)
library(broom)
library(scales)
library(foreign)
library(MASS)
library(brant)
```

1. The Need for an Ordinal Model

We will continue with Regression Analysis on **categorical responses**. Recall that a **discrete response** is considered as categorical if it is of factor-type with more than two categories under the following subclassification:

- **Nominal.** We have categories that do not follow any specific order—for example, the type of dwelling according to the Canadian census: *single-detached house*, *semi-detached house*, *row house*, *apartment*, and *mobile home*.
- **Ordinal.** The categories, in this case, follow a specific order—for example, a Likert scale of survey items: *strongly disagree*, *disagree*, *neutral*, *agree*, and *strongly agree*.

The previous Multinomial Logistic model, from Lecture 2 - Generalized Linear Models: Model Selection and Multinomial Logistic Regression, is unsuitable for ordinal responses. For example, suppose one does not consider the order in the response levels when necessary. In that case, we might risk losing valuable information when making inference or predictions. This valuable information is in the ordinal scale, which could be set up via cumulative probabilities (recall DSCI 551 concepts).

Therefore, the ideal approach is an alternative logistic regression that suits **ordinal** responses. Hence, the existence of the **Ordinal Logistic regression model**. Let us expand the regression mindmap as in <u>Fig. 8</u> to include this new model. Note there is a key characteristic called **proportional odds**, which is reflected in the **data modelling framework** and to be discussed today.



Fig. 8 Expanded regression modelling mind map.

2. College Juniors Dataset

To illustrate the use of this generalized linear model (GLM), let us introduce an adequate dataset.

The College Juniors Dataset

The data frame college_data was obtained from the webpage of the UCLA Statistical Consulting Group.

This dataset contains the results of a survey applied to n=400 college juniors regarding the factors that influence their decision to apply to graduate school.

college_data <- read_csv("../datasets/college_data.csv", show_col_types = FALS
college_data</pre>

Δ	spec	thl	df٠	400	×	3
$\overline{}$	3000	w	uı.	400	\sim	\circ

decision	parent_ed	GPA
<chr></chr>	<chr></chr>	<dbl></dbl>
very likely	No	3.26
somewhat likely	Yes	3.21
unlikely	Yes	3.94
:	÷	÷
somewhat likely	No	2.25
somewhat likely	No	3.26
very likely	No	3.52

The dataset contains the following variables:

- decision: how likely the student will apply to graduate school, a discrete and ordinal response with three increasing categories (unlikely), somewhat likely, and very likely).
- parent_ed: whether at least one parent has a graduate degree, a discrete and binary explanatory variable (Yes and No).

• GPA: the student's current GPA, a continuous explanatory variable.

Important

In this example, variable decision will be our response of interest. However, before proceeding with our regression analysis. Let us set it up as an **ordered factor** via function as ordered().

```
college_data$decision <- as.ordered(college_data$decision)
college_data$decision <- fct_relevel(
   college_data$decision,
   c("unlikely", "somewhat likely", "very likely")
)
levels(college_data$decision)</pre>
```

'unlikely' · 'somewhat likely' · 'very likely'

Main Statistical Inquiries

Let us suppose we want to assess the following:

- Is decision statistically related to parent_ed and GPA?
- How can we interpret these statistical relationships (if there are any!)?

3. Exploratory Data Analysis

Let us plot the data first. Since decision is a **discrete and ordinal response**, we have to be careful about the class of plots we are using regarding exploratory data analysis (EDA).

To explore visually the relationship between decision and GPA, we will use the following side-by-side plots:

- Boxplots.
- Violin plots.

The code below creates side-by-side boxplots and violin plots, where $\overline{\text{GPA}}$ is on the y-axis, and all categories of $\overline{\text{decision}}$ are on the x-axis. The side-by-side violin plots show the mean of $\overline{\text{GPA}}$ by $\overline{\text{decision}}$ category in red points.

```
college data side boxplots <- college data |>
  ggplot(aes(decision, GPA)) +
  geom boxplot(aes(fill = decision)) +
  labs(y = "GPA", x = "Decision to Apply to Graduate School") +
  ggtitle("Side-by-Side Boxplots") +
  theme(
    plot.title = element_text(size = 19, face = "bold"),
    axis.text = element text(size = 17),
    axis.title = element_text(size = 18),
    legend.position = "none"
  ) +
  scale_fill_brewer(palette="BuPu")
college_data_side_violin <- college_data |>
  ggplot(aes(decision, GPA)) +
  geom violin(aes(fill = decision)) +
  labs(y = "GPA", x = "Decision to Apply to Graduate School") +
  ggtitle("Side-by-Side Violin Plots") +
  stat summary(
    fun = mean, colour = "red", geom = "point",
    shape = 18, size = 5
  ) +
  theme(
    plot.title = element_text(size = 19, face = "bold"),
    axis.text = element_text(size = 17),
    axis.title = element text(size = 18),
    legend.position = "none"
  scale fill brewer(palette="BuPu")
```

Now, how can we visually explore the relationship between decision and parent_ed?

In the case of two categorical explanatory variables, we can use stacked bar charts. Each bar will show the percentages (depicted on the y-axis) of college students falling on each ordered category of decision with the categories found in parent_ed on the x-axis. This class of bars will allow us to compare both levels of parent_ed in terms of the categories of decision.

First of all, we need to do some data wrangling. The code below obtains the proportions for each level of decision conditioned on the categories of parent_ed. If we add up all proportions corresponding to No or Yes, we will obtain 1.

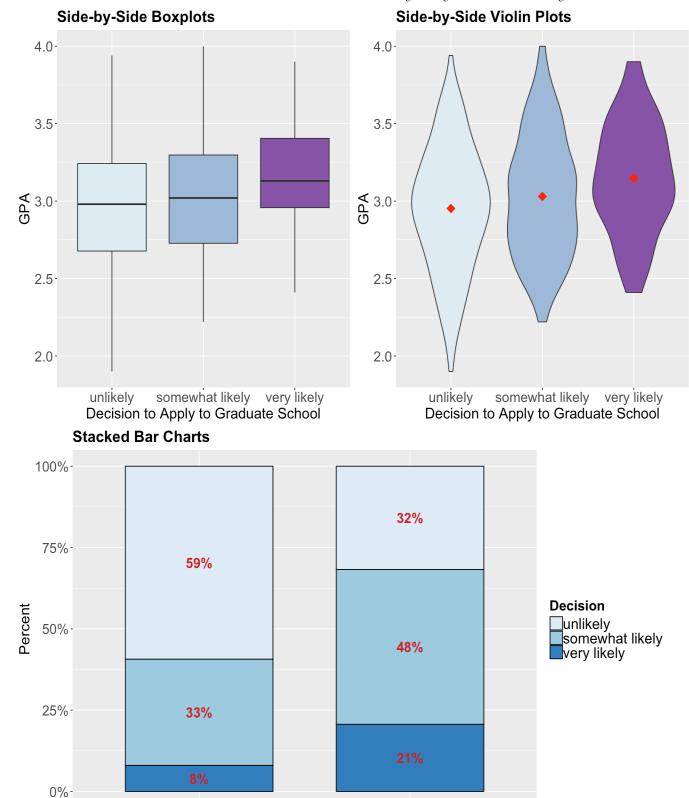
A data.frame: 6×3

parent_ed	decision	prop
<fct></fct>	<fct></fct>	<dbl></dbl>
No	unlikely	0.59347181
Yes	unlikely	0.31746032
No	somewhat likely	0.32640950
Yes	somewhat likely	0.47619048
No	very likely	0.08011869
Yes	very likely	0.20634921

Then, we code the corresponding plot:

```
college data stacked bars <- ggplot(college data prop summary, aes(</pre>
  x = parent ed
  y = prop, fill = decision
  geom_bar(stat = "identity", width = 0.7, colour = "black") +
  geom_text(aes(label = ifelse(prop >= 0.05, paste0(sprintf("%.0f", prop * 100
    position = position_stack(vjust = 0.5), colour = "firebrick3", fontface =
  scale y continuous(labels = percent format()) +
  labs(y = "Percent", x = "Parent Education Status", fill = "") +
  ggtitle("Stacked Bar Charts") +
  theme(
    plot.title = element_text(size = 19, face = "bold"),
    axis.text.x = element text(size = 17, angle = 0),
    axis.text.y = element_text(size = 17, angle = 0),
    axis.title = element text(size = 18),
    legend.text = element_text(size = 18, margin = margin(r = 1, unit = "cm"))
    legend.title = element text(size = 17, face = "bold")
  ) +
  quides(fill = quide_legend(title = "Decision")) +
  scale_fill_brewer(palette = "Blues")
```

```
options(repr.plot.height = 7, repr.plot.width = 12)
plot_grid(college_data_side_boxplots, college_data_side_violin, nrow = 1)
college_data_stacked_bars
```



Yes



Exercise 9

What can we see **descriptively** from the above plots?

Parent Education Status

No



Solution to Exercise 9

Those students with higher GPA tend to be very likely to apply to graduate school. Interestingly, those students with at least one parent with a graduate degree tend to be very likely to apply to graduate school.

4. Data Modelling Framework

Let us suppose that a given **discrete ordinal response** Y_i (for $i=1,\ldots,n$) has categories $1, 2, \ldots, m$ in a training set of size n.

Important

Categories $1, 2, \ldots, m$ implicate an ordinal scale here, i.e., $1 < 2 < \cdots < m$.

Also, note there is more than one class of Ordinal Logistic regression. We will review the proportional odds version, which is also a cumulative logit model.

We set up the Ordinal Logistical regression model for this specific dataset with m=3. For the *i*th observation with the continuous $X_{i,\mathtt{GPA}}$ and the dummy variable

$$X_{i,\mathtt{parent_ed}} = egin{cases} 1 & ext{at least one parent has a graduate degree}, \ 0 & ext{neither of the parents has a graduate degree}; \end{cases}$$

the model will indicate how each one of the regressors affects the cumulative logarithm of **the odds** in [decision] for the following 2 situations:

Level somewhat likely or any lesser degree versus level very likely Level unlikely versus level somewhat likely or any higher degree

Then, we have the following system of 2 equations:

$$\begin{split} \eta_i^{(\texttt{somewhat likely})} &= \log \left[\frac{P(Y_i \leq \texttt{somewhat likely} \mid X_{i,\texttt{GPA}}, X_{i,\texttt{parent_ed}})}{P(Y_i = \texttt{very likely} \mid X_{i,\texttt{GPA}}, X_{i,\texttt{parent_ed}})} \right] \\ &= \beta_0^{(\texttt{somewhat likely})} - \beta_1 X_{i,\texttt{GPA}} - \beta_2 X_{i,\texttt{parent_ed}} \end{split}$$

$$\eta_i^{(ext{unlikely})} = \log \left[rac{P(Y_i = ext{unlikely} \mid X_{i, ext{GPA}}, X_{i, ext{parent_ed}})}{P(Y_i \geq ext{somewhat likely} \mid X_{i, ext{GPA}}, X_{i, ext{parent_ed}})}
ight] \ = eta_0^{(ext{unlikely})} - eta_1 X_{i, ext{GPA}} - eta_2 X_{i, ext{parent_ed}}.$$

The system has 2 intercepts but only 2 regression coefficients. Check the signs of the common β_1 and β_2 , which are minuses. This is a fundamental parameterization of the proportional odds model.

To make coefficient interretation easier, we could re-express the equations of $\eta_i^{(\text{somewhat likely})}$ and $\eta_i^{(\text{unlikely})}$ as follows:

$$\frac{P(Y_i = \mathtt{very\ likely} \mid X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})}{P(Y_i \leq \mathtt{somewhat\ likely} \mid X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})} = \exp\big(-\beta_0^{(\mathtt{somewhat\ likely})}\big) \exp\big(\beta_1$$

$$\frac{P(Y_i \geq \mathtt{somewhat\ likely} \mid X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})}{P(Y_i = \mathtt{unlikely} \mid X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})} = \exp\big(-\beta_0^{(\mathtt{unlikely})}\big) \exp\big(\beta_1 X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}}\big)$$

Both regression coefficients (β_1 and β_2) have a multiplicative effect on the odds on the left-hand side. Moreover, **given the proportional odds assumptions**, the regressor effects are the same for both cumulative odds.

What is the math for the general case with m response categories and k regressors?

Before going into the model's mathematical notation, we have to point out that Ordinal Logistic regression will indicate how each one of the k regressors $X_{i,1},\ldots,X_{i,k}$ affects the **cumulative logarithm of the odds** in the ordinal response for the following m-1 situations:

Level m-1 or any lesser degree versus level mLevel m-2 or any lesser degree versus level m-1 or any higher degree \vdots

Level 2 or any lesser degree versus level 3 or any higher degree Level 1 versus level 2 or any higher degree

These m-1 situations are translated into cumulative probabilities using the logarithms of the odds on the left-hand side (m-1) link functions subject to the linear combination of the k regressors $X_{i,j}$ (for $j=1,\ldots,k$):

$$\begin{split} \eta_i^{(m-1)} &= \log \left[\frac{P(Y_i \leq m-1 \mid X_{i,1}, \dots, X_{i,k})}{P(Y_i = m \mid X_{i,1}, \dots, X_{i,k})} \right] = \beta_0^{(m-1)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \dots - \beta_n X_{i,n} \\ \eta_i^{(m-2)} &= \log \left[\frac{P(Y_i \leq m-2 \mid X_{i,1}, \dots, X_{i,k})}{P(Y_i > m-2 \mid X_{i,1}, \dots, X_{i,k})} \right] = \beta_0^{(m-2)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \dots - \beta_n X_{i,n} \\ &\vdots \\ \eta_i^{(2)} &= \log \left[\frac{P(Y_i \leq 2 \mid X_{i,1}, \dots, X_{i,k})}{P(Y_i > 2 \mid X_{i,1}, \dots, X_{i,k})} \right] = \beta_0^{(2)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \dots - \beta_n X_{i,k} \\ \eta_i^{(1)} &= \log \left[\frac{P(Y_i = 1 \mid X_{i,1}, \dots, X_{i,k})}{P(Y_i > 1 \mid X_{i,1}, \dots, X_{i,k})} \right] = \beta_0^{(1)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \dots - \beta_n X_{i,k} \end{split}$$

Note that the system above has m-1 intercepts but **only** k **regression coefficients**. In general, the previous m-1 equations can be generalized for levels $j=m-1,\ldots,1$ as follows:

$$egin{aligned} \eta_i^{(j)} &= \log \left[rac{P(Y_i \leq j \mid X_{i,1}, \dots, X_{i,k})}{P(Y_i > j \mid X_{i,1}, \dots, X_{i,k})}
ight] = eta_0^{(j)} - eta_1 X_{i,1} - eta_2 X_{i,2} - \dots - eta_k X_{i,k} \ &\Rightarrow P(Y_i \leq j \mid X_{i,1}, \dots, X_{i,k}) = rac{\exp \left(eta_0^{(j)} - eta_1 X_{i,1} - eta_2 X_{i,2} - \dots - eta_k X_{i,k}
ight)}{1 + \exp \left(eta_0^{(j)} - eta_1 X_{i,1} - eta_2 X_{i,2} - \dots - eta_k X_{i,k}
ight)} \end{aligned}$$

The probability that Y_i will fall in the category j can be computed as follows:

$$p_{i,j} = P(Y_i = j \mid X_{i,1}, \dots, X_{i,k}) = P(Y_i \leq j \mid X_{i,1}, \dots, X_{i,k}) - P(Y_i \leq j-1 \mid X_{i,1}, \dots)$$

which leads to

$$P(Y_i = 1) = p_{i,1}$$
 $P(Y_i = 2) = p_{i,2}$... $P(Y_i = m) = p_{i,m}$

where

$$\sum_{j=1}^m p_{i,j} = p_{i,1} + p_{i,2} + \cdots + p_{i,m} = 1.$$



Attention

You will prove why we have **proportional odds** in this model in the challenging **Exercise 4** of lab2.

Let us check the next in-class question via iClicker.



Exercise 10

Suppose we have a case with an ordered response of five levels and four continuous regressors; how many link functions, intercepts, and coefficients will our Ordinal Logistic regression model have under a proportional odds assumption?

- **A.** Four link functions, four intercepts, and sixteen regression coefficients.
- **B.** Five link functions, five intercepts, and five regression coefficients.
- **C.** Five link functions, five intercepts, and twenty regression coefficients.
- **D.** Four link functions, four intercepts, and four regression coefficients.

Solution to Exercise 10

The categories 1, 2, 3, 4, 5 in the ith response Y_i implicate an ordinal scale here, i.e., 1 < 2 < 3 < 4 < 5.

Hence, our Ordinal Logistic regression model will indicate how each one of the 4 regressors $X_{i,1}$, $X_{i,2}$, $X_{i,3}$, and $X_{i,4}$ affects the **cumulative logarithm of the odds** in the ordinal response for the following 4 situations:

Level 4 or any lesser degree versus level 5 Level 3 or any lesser degree versus level 4 or any higher degree Level 2 or any lesser degree versus level 3 or any higher degree Level 1 versus level 2 or any higher degree

Then, our system of modelling equations will be:

$$\eta_{i}^{(4)} = \log \left[\frac{P(Y_{i} \leq 4 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} = 5 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(4)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(3)} = \log \left[\frac{P(Y_{i} \leq 3 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 3 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(3)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(2)} = \log \left[\frac{P(Y_{i} \leq 2 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 2 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(2)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right] = \beta_{0}^{(1)} - \beta_{1}X_{i,1} - \beta_{2}X_{i,2} - \beta_{3}X_{i,3} - \eta_{i}^{(1)} = \log \left[\frac{P(Y_{i} = 1 \mid X_{i,1}, \dots, X_{i,4})}{P(Y_{i} > 1 \mid X_{i,1}, \dots, X_{i,4})} \right]$$

As we can see above; we would need four link functions, four intercepts $\left(\beta_0^{(4)},\beta_0^{(3)},\beta_0^{(2)},\beta_1^{(1)}\right)$, and four coefficients $(\beta_1,\beta_2,\beta_3,\beta_4)$.

A Graphical Intuition of the Proportional Odds Assumption

The challenging **Exercise 4** in Tab2 is about showing mathematically what the proportional odds assumption stands for in the Ordinal Logistic regression model with k regressors and a training set size of n, using a response of m levels in general. That said, let us show the graphical intuition with an example of an ordinal response of 5 levels and a single continuous regressor $X_{i,1}$ (which is unbounded!).

The system of cumulative logit equations is the following:

$$egin{aligned} \eta_i^{(4)} &= \log \left[rac{P(Y_i \leq 4 \mid X_{i,1})}{P(Y_i = 5 \mid X_{i,1})}
ight] = eta_0^{(4)} - eta_1 X_{i,1} \ \eta_i^{(3)} &= \log \left[rac{P(Y_i \leq 3 \mid X_{i,1})}{P(Y_i > 3 \mid X_{i,1})}
ight] = eta_0^{(3)} - eta_1 X_{i,1} \ \eta_i^{(2)} &= \log \left[rac{P(Y_i \leq 2 \mid X_{i,1})}{P(Y_i > 2 \mid X_{i,1})}
ight] = eta_0^{(2)} - eta_1 X_{i,1} \ \eta_i^{(1)} &= \log \left[rac{P(Y_i = 1 \mid X_{i,1})}{P(Y_i > 1 \mid X_{i,1})}
ight] = eta_0^{(1)} - eta_1 X_{i,1}. \end{aligned}$$

This above system could be mathematically expressed as follows:

$$egin{aligned} P(Y_i \leq 4 \mid X_{i,1}) &= rac{\exp\left(eta_0^{(4)} - eta_1 X_{i,1}
ight)}{1 + \exp\left(eta_0^{(4)} - eta_1 X_{i,1}
ight)} \ P(Y_i \leq 3 \mid X_{i,1}) &= rac{\exp\left(eta_0^{(3)} - eta_1 X_{i,1}
ight)}{1 + \exp\left(eta_0^{(3)} - eta_1 X_{i,1}
ight)} \ P(Y_i \leq 2 \mid X_{i,1}) &= rac{\exp\left(eta_0^{(2)} - eta_1 X_{i,1}
ight)}{1 + \exp\left(eta_0^{(2)} - eta_1 X_{i,1}
ight)} \ P(Y_i \leq 1 \mid X_{i,1}) &= rac{\exp\left(eta_0^{(1)} - eta_1 X_{i,1}
ight)}{1 + \exp\left(eta_0^{(1)} - eta_1 X_{i,1}
ight)}. \end{aligned}$$

Note we have four intercepts $(\beta_0^{(1)}, \beta_0^{(2)}, \beta_0^{(3)}, \beta_0^{(4)})$ and a single coefficient β_1 . Moreover, let us assume we estimate these five parameters to be the following:

$$egin{aligned} \hat{eta}_0^{(1)} &= 4 \ \hat{eta}_0^{(2)} &= 5.5 \ \hat{eta}_0^{(3)} &= 6.3 \ \hat{eta}_0^{(4)} &= 10 \ \hat{eta}_1 &= 12. \end{aligned}$$

Let us check our estimated sigmoid functions! As we can see in <u>Fig. 9</u>, the proportional odds assumption also implicates proportional sigmoid functions.

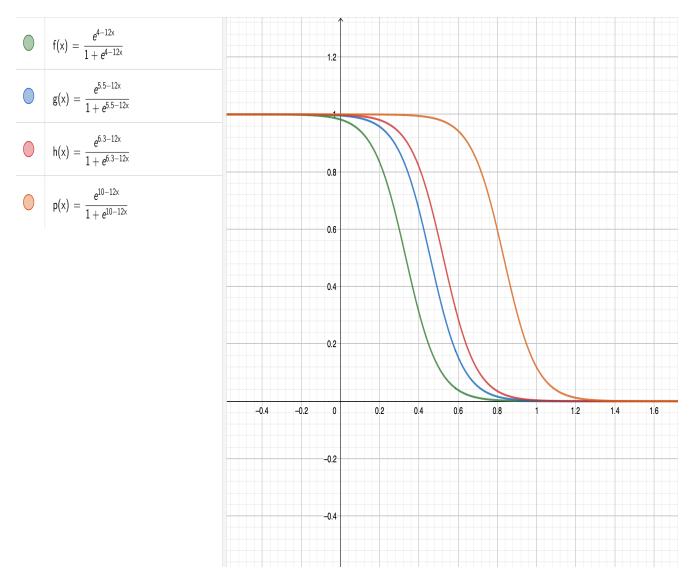


Fig. 9 Sigmoid functions under a proportional odds assumption (Copyright GeoGebra®, 2024).

5. Estimation

All parameters in the Ordinal Logistic regression model are also unknown. Therefore, model estimates are obtained through **maximum likelihood**, where we also assume a **Multinomial joint probability mass function** of the n responses Y_i . Moreover, this Multinomial assumption plays around with cumulative probabilities in the joint likelihood function **as discussed in Agresti (2013) in Chapter 8 (Section 8.2.2)**.

To fit the model with the package MASS, we use the function polr(), which obtains the corresponding estimates. The argument Hess = TRUE is required to compute the Hessian matrix of the log-likelihood function, which is used to obtain the standard errors of the estimates.

```
ordinal_model <- polr(decision ~ parent_ed + GPA,
  data = college_data, Hess = TRUE
)</pre>
```

Important

Maximum likelihood estimation in this regression model is quite complex compared to the challenging **Exercise 5** in lab1 for Binary Logistic regression. Thus, **the proof will be out of the scope of this course.**

Note

If you would like to check a further rationale on the use of the Hessian matrix to obtain the standard errors of regression estimates, you can find good insights into this matter via the classical linear regression model (i.e., Ordinary Least-squares but in its maximum likelihood estimation flavour) in this resource.

6. Inference

We can determine whether a regressor is statistically associated with the logarithm of the cumulative odds through hypothesis testing for the parameters β_j . We also use the Wald statistic z_j :

$$z_j = rac{\hat{eta}_j}{\mathrm{se}(\hat{eta}_j)}$$

to test the hypotheses

$$H_0: eta_j = 0$$

 $H_a: eta_i
eq 0.$

The **null hypothesis** H_0 indicates that the jth regressor associated to β_j does not affect the response variable in the model, and the **alternative hypothesis** H_a otherwise. Moreover, provided the sample size n is large enough, z_j has an **approximately Standard Normal distribution** under H_0 .

R provides the corresponding p-values for each β_j . The smaller the p-value, the stronger the evidence against the null hypothesis H_0 . As in the previous regression models, we would set a predetermined significance level α (usually taken to be 0.05) to infer if the p-value is small enough. If the p-value is smaller than the predetermined level α , then you could claim that there is evidence to reject the null hypothesis. Hence, p-values that are small enough indicate that the data provides evidence in favour of association (or causation in the case of an experimental study!) between the response variable and the jth regressor.

Furthermore, given a specified level of confidence where α is the significance level, we can construct approximate $(1-\alpha)\times 100\%$ confidence intervals for the corresponding true value of β_i :

$$\hat{eta}_j \pm z_{lpha/2} {
m se}(\hat{eta}_j),$$

where $z_{lpha/2}$ is the upper lpha/2 quantile of the **Standard Normal distribution**.

Function [polr()] does not provide p-values, but we can compute them as follows:

```
summary_ordinal_model <- cbind(tidy(ordinal_model),
  p.value = pnorm(abs(tidy(ordinal_model)$statistic),
    lower.tail = FALSE
  ) * 2
) |>
    mutate_if(is.numeric, round, 2)
summary_ordinal_model
```

A data.frame: 4×6

term	estimate	std.error	statistic	coef.type	p.value
<chr></chr>	<dbl></dbl>	<dbl></dbl>	<dbl></dbl>	<chr></chr>	<dbl></dbl>
parent_edYes	1.05	0.27	3.94	coefficient	0.00
GPA	0.60	0.25	2.38	coefficient	0.02
unlikely somewhat likely	2.18	0.77	2.84	scale	0.00
somewhat likely very likely	4.27	0.79	5.39	scale	0.00

Terms [unlikely|somewhat likely] and [somewhat likely|very likely] refer to the estimated intercepts $\hat{\beta}_0^{(\text{unlikely})}$ and $\hat{\beta}_0^{(\text{somewhat likely})}$, respectively (note the hat notation).

Finally, we see that those coefficients associated with GPA and parent_ed are statistically significant with lpha=0.05.

Function confint() can provide the 95% confidence intervals for the estimates contained in tidy().

round(confint(ordinal_model), 2)

Waiting for profiling to be done...

A matrix: 2×2 of type dbl

	2.5 %	97.5 %
parent_edYes	0.53	1.57
GPA	0.11	1.11

7. Coefficient Interpretation

The interpretation of the Ordinal Logistic regression coefficients will change since we are modelling cumulative probabilities. Let us interpret the association of decision with GPA and parent_ed.

tibble(summary_ordinal_model[1:2, 1:2], exp.estimate = round(exp(summary_ordin

A tibble: 2×3

term	estimate	exp.estimate	
<chr></chr>	<dbl></dbl>	<dbl></dbl>	
parent_edYes	1.05	2.86	
GPA	0.60	1.82	

By using the column exp.estimate, along with the model equations on the original scale of the cumulative odds, we interpret the two regression coefficients above by each odds as follows:

For the odds

$$\frac{P(Y_i = \mathtt{very\ likely} \mid X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})}{P(Y_i \leq \mathtt{somewhat\ likely} \mid X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})} = \exp\big(-\beta_0^{(\mathtt{somewhat\ likely})}\big) \exp\big(\beta_1 + \beta_0^{(\mathtt{somewhat\ likely})}\big)$$

- eta_1 : "for each one-unit increase in the GPA, the odds that the student is very likely versus somewhat likely or unlikely to apply to graduate school increase by $\exp\left(\hat{eta}_1\right)=1.82$ times (while holding parent_ed constant)."
- eta_2 : "for those respondents whose parents attended to graduate school, the odds that the student is very likely versus somewhat likely or unlikely to apply to graduate school increase by $\exp\left(\hat{eta}_2\right)=2.86$ times (when compared to those respondents whose parents did not attend to graduate school and holding GPA constant)."

For the odds

$$\frac{P(Y_i \geq \mathtt{somewhat\ likely} \mid X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})}{P(Y_i = \mathtt{unlikely} \mid X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})} = \exp\big(-\beta_0^{(\mathtt{unlikely})}\big) \exp\big(\beta_1 X_{i,\mathtt{GPA}}, X_{i,\mathtt{parent_ed}})$$

- eta_1 : "for each one-unit increase in the GPA, the odds that the student is very likely or somewhat likely versus unlikely to apply to graduate school increase by $\exp\left(\hat{eta}_1\right)=1.82$ times (while holding parent_ed constant)."
- eta_2 : "for those respondents whose parents attended to graduate school, the odds that the student is very likely or somewhat likely versus unlikely to apply to graduate school increase by $\exp\left(\hat{eta}_2\right)=2.86$ times (when compared to those respondents whose parents did not attend to graduate school and holding GPA constant)."

8. Predictions

Aside from our main statistical inquiries, using the function predict() with the object ordinal_model, we can obtain the estimated probabilities to apply to graduate school (associated to unlikely, somewhat likely, and very likely) for a student with a GPA of 3.5 whose parents attended to graduate school.

```
round(predict(ordinal_model, tibble(GPA = 3.5, parent_ed = "Yes"), type = "pro
```

unlikely: 0.27 somewhat likely: 0.48 very likely: 0.25

We can see that it is somewhat likely to apply to graduate school with a probability of 0.48. Thus, we could classify the student in this ordinal category.

Unfortunately, <code>predict()</code> with <code>polr()</code> does not provide a quick way to compute this model's corresponding predicted cumulative odds for a new observation. Nonetheless, we could use function <code>vglm()</code> (from package <code>VGAM</code>) again while keeping in mind that **this**

function labels the response categories as numbers. Let us check this for decision with levels().

```
levels(college_data$decision)
```

'unlikely' · 'somewhat likely' · 'very likely'

Hence, unlikely is 1, somewhat likely is 2, and very likely is 3. Now, let us use vglm() with propodds to fit the same model (named ordinal_model_vglm) as follows:

```
ordinal_model_vglm <- vglm(decision ~ parent_ed + GPA, propodds,
   data = college_data)
summary(ordinal_model_vglm)</pre>
```

```
Call:
vglm(formula = decision ~ parent_ed + GPA, family = propodds,
    data = college_data)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -2.1763
                           0.7726 - 2.817 0.00485 **
(Intercept):2 -4.2716
                           0.7976 -5.356 8.52e-08 ***
parent edYes
                1.0457
                           0.2682 3.899 9.67e-05 ***
GPA
                0.6043
                           0.2561
                                  2.360 0.01829 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])
Residual deviance: 717.0638 on 796 degrees of freedom
Log-likelihood: -358.5319 on 796 degrees of freedom
Number of Fisher scoring iterations: 4
No Hauck-Donner effect found in any of the estimates
Exponentiated coefficients:
parent edYes
                      GPA
    2.845388
                 1.829895
```

We will double-check whether ordinal_model_vglm provides the same classification probabilities via predict(). Note that the function parameterization changes to <a href="type="type="type-style-typ

round(predict(ordinal_model_vglm, tibble(GPA = 3.5, parent_ed = "Yes"), type =

A matrix: 1×3 of type dbl

	unlikely	somewhat likely	very likely
1	0.27	0.48	0.25

Yes, we get identical results using vglm(). Now, let us use predict() with $ordinal_model_vglm$ to obtain the predicted cumulative odds for a student with a vglm() of 3.5 whose parents attended to graduate school (via argument vglm() in vglm()).

```
round(predict(ordinal_model_vglm, tibble(GPA = 3.5, parent_ed = "Yes"), type =
```

A matrix: 1×2 of type dbl

	logitlink(P[Y>=2])	logitlink(P[Y>=3])
1	0.98	-1.11

Then, let us take these predictions to the odds' original scale via exp().

```
round(exp(predict(ordinal_model_vglm, tibble(GPA = 3.5, parent_ed = "Yes"), ty
```

A matrix: 1×2 of type dbl

	logitlink(P[Y>=2])	logitlink(P[Y>=3])
1	2.68	0.33

Again, we see the numeric labels (a drawback of using vglm()). We interpret these predictions as follows:

• $\frac{P(Y_i = \text{very likely} | X_{i,\text{GPA}}, X_{i,\text{parent_ed}})}{P(Y_i \leq \text{somewhat likely} | X_{i,\text{GPA}}, X_{i,\text{parent_ed}})}$, i.e., P[Y>=3] (label 3, very likely): "a student, with a GPA of 3.5 whose parents attended to graduate school, is 3.03 (1/0.33) times somewhat likely or unlikely versus very likely to apply to graduate school." This prediction goes in line with the predicted probabilities above,

given that somewhat likely has a probability of 0.48 and unlikely a probability of 0.27.

• $\frac{P(Y_i \geq \text{somewhat likely}|X_{i,\text{GPA}},X_{i,\text{parent_ed}})}{P(Y_i = \text{unlikely}|X_{i,\text{GPA}},X_{i,\text{parent_ed}})}, \text{ i.e., } P[Y>=2] \text{ (label 2, somewhat likely): "a student, with a GPA of 3.5 whose parents attended to graduate school, is 2.68 times very likely or somewhat likely versus unlikely to apply to graduate school." This prediction goes in line with the predicted probabilities above, given that somewhat likely has a probability of 0.48 and very likely a probability of 0.25.$

9. Model Selection

To perform model selection, we can use the same techniques from the Binary Logistic regression model (check <u>8. Model Selection</u>). That said, some <u>R</u> coding functions might not be entirely available for the <u>polr()</u> models. Still, these statistical techniques and metrics can be manually coded.

10. Non-proportional Odds

Now, we might wonder

What happens if we do not fulfil the proportional odds assumption in our Ordinal Logistic regression model?

10.1. The Brant-Wald Test

It is essential to remember that the Ordinal Logistic model under the proportional odds assumption is **the first step** when performing Regression Analysis on an ordinal response.

Once this model has been fitted, it is possible to assess whether it fulfils this strong assumption statistically.

We can do it via the Brant-Wald test:

• This tool statistically assesses whether our model globally fulfils this assumption. It has the following hypotheses:

 H_0 : Our Ordinal Logistic regression model globally fulfils the proportional odds as H_a : Otherwise.

• Moreover, it also performs further hypothesis testing on each regressor. With k regressors for $j=1,\ldots,k$; we have the following hypotheses:

 H_0 : The jth regressor in our Ordinal Logistic regression model fulfils the proportio H_a : Otherwise.

Function brant() from package brant implements this tool, which can be used in our polr() object ordinal_model:

brant(ordinal_model)

Test for	X2	df	probability
Omnibus parent_edYes GPA	0.73	2	0.69
	0.09	1	0.76
	0.7	1	0.4

HO: Parallel Regression Assumption holds

The row <code>Omnibus</code> represents the global model, while the other two rows correspond to our two regressors: <code>parent_ed</code> and <code>GPA</code>. Note that with $\alpha=0.05$, we are completely fulfilling the proportional odds assumption (the column <code>probability</code> delivers the corresponding p-values).

Important

The Brant Wald test essentially compares this basic Ordinal Logistic regression model of k-1 cumulative logit functions versus a combination of k-1 correlated Binary Logistic regressions.

10.2. A Non-proportional Odds Model

Suppose that our example case does not fulfil the proportional odds assumption according to the Brant Wald test. It is possible to have a model under a non-proportional odds assumption as follows:

$$\begin{split} \eta_i^{(\texttt{somewhat likely})} &= \log \left[\frac{P(Y_i \leq \texttt{somewhat likely} \mid X_{i,\texttt{GPA}}, X_{i,\texttt{parent_ed}})}{P(Y_i = \texttt{very likely} \mid X_{i,\texttt{GPA}}, X_{i,\texttt{parent_ed}})} \right] \\ &= \beta_0^{(\texttt{somewhat likely})} - \beta_1^{(\texttt{somewhat likely})} X_{i,\texttt{GPA}} - \beta_2^{(\texttt{somewhat likely})} X_{i,\texttt{parent_ed}} \end{split}$$

$$\eta_i^{(ext{unlikely})} = \log \left[rac{P(Y_i = ext{unlikely} \mid X_{i, ext{GPA}}, X_{i, ext{parent_ed}})}{P(Y_i \geq ext{somewhat likely} \mid X_{i, ext{GPA}}, X_{i, ext{parent_ed}})}
ight] \ = eta_0^{(ext{unlikely})} - eta_1^{(ext{unlikely})} X_{i, ext{GPA}} - eta_2^{(ext{unlikely})} X_{i, ext{parent_ed}}.$$

This is called a Generalized Ordinal Logistic regression model. This class of models can be fit via the function cumulative() from package (VGAM). Note that estimation, inference, coefficient interpretation, and predictions are conducted in a similar way compared to the proportional odds model.



Exercise 11

Why would we have to be cautious when fitting a non-proportional odds model in the context of a response with a considerable number of ordinal levels and regressors?



Solution to Exercise 11

The Ordinal Logistic regression framework makes complex modelling since we deal with cumulative probabilities by each logit function in the system. On top of that, adding a considerable number of regression parameters to estimate in our optimization procedure via maximum likelihood might make convergence slow when fitting the model.

11. Wrapping Up on Categorical Regression Models

- Multinomial and Ordinal Logistic regression models address cases where our response is discrete and categorical. These models are another class GLMs with their corresponding link functions.
- Multinomial Logistic regression approaches nominal responses. It heavily relies on a baseline response level and the odds' natural logarithm.
- Ordinal Logistic regression approaches ordinal responses. It relies on cumulative odds on an ordinal scale.
- We have to be careful with coefficient interpretations in these models, especially with the cumulative odds. This will imply an effective communication of these estimated models for inferential matters.