STA 610L: Module 4.9

Multilevel categorical outcomes

Dr. Olanrewaju Michael Akande

Categorical data

We've focused on hierarchical models for binary and continuous data.

Of course, our data may follow a wide variety of distributions.

Today we'll consider extensions to categorical data, as interpretations of these models may be less straightforward than extensions to say count data.

Examples of categorical data: beverage order in a restaurant (water, tea, coffee, soda, wine, beer, mixed drink) or your favorite Duke stats professor.

First we will review simple logistic regression, and then extend the ideas to multiple outcomes.

Recall logistic regression

Recall that for the simple logistic regression model, we had

$$y_i|x_i \sim \text{Bernoulli}(\pi_i); \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

for each observation $i = 1, \ldots, n$.

To get π_i, we solved the logit equation above to get

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

Consider $Y = 0$ a baseline category. Suppose $\Pr[y_i = 1|x_i] = \pi_{i1}$ and $\Pr[y_i = 0|x_i] = \pi_{i0}$. Then, the logit expression is essentially

$$\log\left(\frac{\pi_{i1}}{\pi_{i0}}\right) = \beta_0 + \beta_1 x_i.$$

e^{β_1} is thus the (multiplicative) change in odds of $y = 1$ over the baseline $y = 0$ when increasing x by one unit.

Multinomial logistic regression

Suppose we have a nominal-scale response variable Y with K categories, that is, $Y = 1, \ldots, K$.

First, for the random component, we need a distribution to describe Y.

A standard option for this is the multinomial distribution. THe distribution gives us a way to characterize

$$\Pr[y_i = 1] = \pi_1,\ Pr[y_i = 2] = \pi_2,\ \ldots,\ \Pr[y_i = K] = \pi_K, \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1.$$

When there are no predictors, the best guess for each π_k is the sample proportion of cases with $y_i = k$, that is,

$$\hat{\pi}_k = \frac{\mathbf{1}[y_i = k]}{n}.$$

When we have predictors, then we want

$$\Pr[y_i = 1|\boldsymbol{x}_i] = \pi_{i1},\ \Pr[y_i = 2|\boldsymbol{x}_i] = \pi_{i2},\ \ldots,\ \Pr[y_i = K|\boldsymbol{x}_i] = \pi_{iK}.$$

Multinomial logistic regression

That is, we want the π_k's to be functions of the predictors, like in logistic regression.

Turns out we can use the same link function, that is the logit function, if we set one of the levels as the baseline.

Pick a baseline outcome level, say $Y = 1$.

Then, the multinomial logistic regression is defined as a set of logistic regression models for each probability π_k, compared to the baseline, where $k \geq 2$.

That is,

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \ldots + \beta_{pk}x_{ip}; \quad k \geq 2.$$

We therefore have $K - 1$ separate logistic regressions in this setup.

Multinomial logistic regression

The equation for each π_{ik} is given by

$$\pi_{ik} = \frac{e^{\beta_{0k}+\beta_{1k}x_{i1}+\beta_{2k}x_{i2}+\ldots+\beta_{pk}x_{ip}}}{1+\sum_{k=2}^{K} e^{\beta_{0k}+\beta_{1k}x_{i1}+\beta_{2k}x_{i2}+\ldots+\beta_{pk}x_{ip}}} \quad \text{for} \quad k \geq 2$$

and

$$\pi_{i1} = 1 - \sum_{k=2}^{K} \pi_{ik}.$$

Also, we can extract the log odds for comparing other pairs of the response categories k and $k^\star$, since

$$\begin{aligned}
\log\left(\frac{\pi_{ik}}{\pi_{ik^\star}}\right) &= \log\left(\pi_{ik}\right) - \log\left(\pi_{ik^\star}\right) \\
&= \log\left(\pi_{ik}\right) - \log\left(\pi_{i1}\right) - \log\left(\pi_{ik^\star}\right) + \log\left(\pi_{i1}\right) \\
&= \left[\log\left(\pi_{ik}\right) - \log\left(\pi_{i1}\right)\right] - \left[\log\left(\pi_{ik^\star}\right) - \log\left(\pi_{i1}\right)\right] \\
&= \log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) - \log\left(\frac{\pi_{ik^\star}}{\pi_{i1}}\right).
\end{aligned}$$

Multinomial logistic regression

Each coefficient has to be interpreted relative to the baseline.

Each β_{0k} represents the baseline log-odds of general preference for $Y = k$ over $Y = 1$.

That is, for a continuous predictor,

- β_{1k} is the increase (or decrease) in the log-odds of $Y = k$ versus $Y = 1$ when increasing x_1 by one unit.
- $e^{\beta_{1k}}$ is the multiplicative increase (or decrease) in the odds of $Y = k$ versus $Y = 1$ when increasing x_1 by one unit.

Exponentiate confidence intervals from log-odds scale to get on the odds scale.

Multinomial logistic regression

Whereas, for a binary predictor,

- β_{1k} is the log-odds of $Y = k$ versus $Y = 1$ for the group with $x_1 = 1$, compared to the group with $x_1 = 0$.
- $e^{\beta_{1k}}$ is the odds of $Y = k$ versus $Y = 1$ for the group with $x_1 = 1$, compared to the group with $x_1 = 0$.

Again, exponentiate confidence intervals from log-odds scale to get on the odds scale.

Model diagnostics

Use binned residuals like in logistic regression.

Each outcome level has its own raw residual. For each outcome level k,

- make an indicator variable equal to one whenever $Y = k$ and equal to zero otherwise;
- compute the predicted probability that $Y = k$ for each record; and
- compute the raw residual = indicator value - predicted probability.

For each outcome level, make bins of predictor values and plot average value of predictor versus the average raw residual. Look for patterns.

You can still compute accuracy just as in the logistic regression model.

ROC on the other hand is not so straightforward; we can draw a different ROC curve for each level of the response variable. We can also draw pairwise ROC curves.

Hierarchical extension

Consider the model:

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \beta_{0k} + \beta_{1k}x_i; \quad k \geq 2.$$

Suppose we now have multiple measurements j per participant i in a study or per group.

For example, we might ask about instructor preference for a list of courses.

How might we add random effects to this model?

Hierarchical extension

You don't want to assume that just because a participant has more of a tendency to select category 2 than category 1, they will also have more of a tendency to select category 3 than category 1.

Thus a single random intercept per person may be insufficient.

We want to allow $k-1$ random intercepts per person.

That is,

$$\log\left(\frac{\pi_{ijk}}{\pi_{ij1}}\right) = \beta_{0k} + \beta_{1k}x_{ij} + b_{ik}; \quad k \geq 2, \;\; b_{ik} \sim N(0, \sigma_k^2).$$

Example: clarity of inhaler instructions

Ezzet and Whitehead (1991) present data from an industry-sponsored clinical trial designed to evaluate the clarity of two different sets of instructions for using two different inhalers (the variable treat indicates the inhaler used and is coded 0.5 and -0.5) to deliver an asthma drug.

Each participant rated each inhaler; the variable period indicates whether the rating is from the first or second inhaler evaluated (in case participants learned from the first evaluation).

The order of evaluation was randomized across subjects.

After using a device, the participant rated (variable name: rating) the instruction leaflet as

- 1 = easy to understand;
- 2 = only clear after rereading;
- 3 = not very clear;
- 4 = confusing.

Clarity of inhaler instructions

```
data(inhaler); head(inhaler)
```

```
##   subject rating treat period carry
## 1       1      1   0.5    0.5     0
## 2       2      1   0.5    0.5     0
## 3       3      1   0.5    0.5     0
## 4       4      1   0.5    0.5     0
## 5       5      1   0.5    0.5     0
## 6       6      1   0.5    0.5     0
```

```
#note, carry variable is a contrast to indicate possible carry over effects
#we won't use the variable
inhaler$treat <- as.factor(inhaler$treat)
inhaler$period <- as.factor(inhaler$period)
inhaler$rating <- as.ordered(inhaler$rating)
table(inhaler$treat)
```

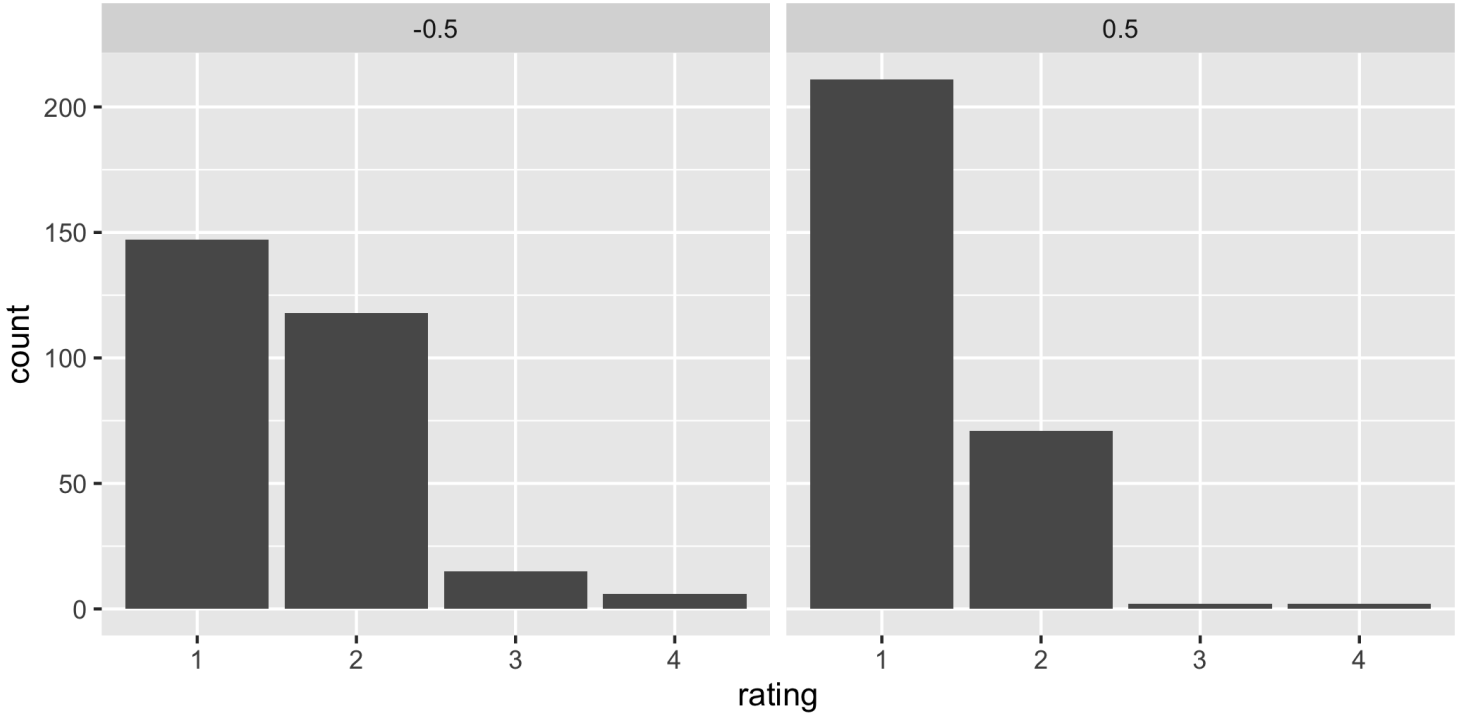
```
##
## -0.5  0.5
##  286  286
```

```
table(inhaler$treat, inhaler$period)
```

```
##
##        -0.5 0.5
##   -0.5  142 144
##   0.5   144 142
```

Clarity of inhaler instructions

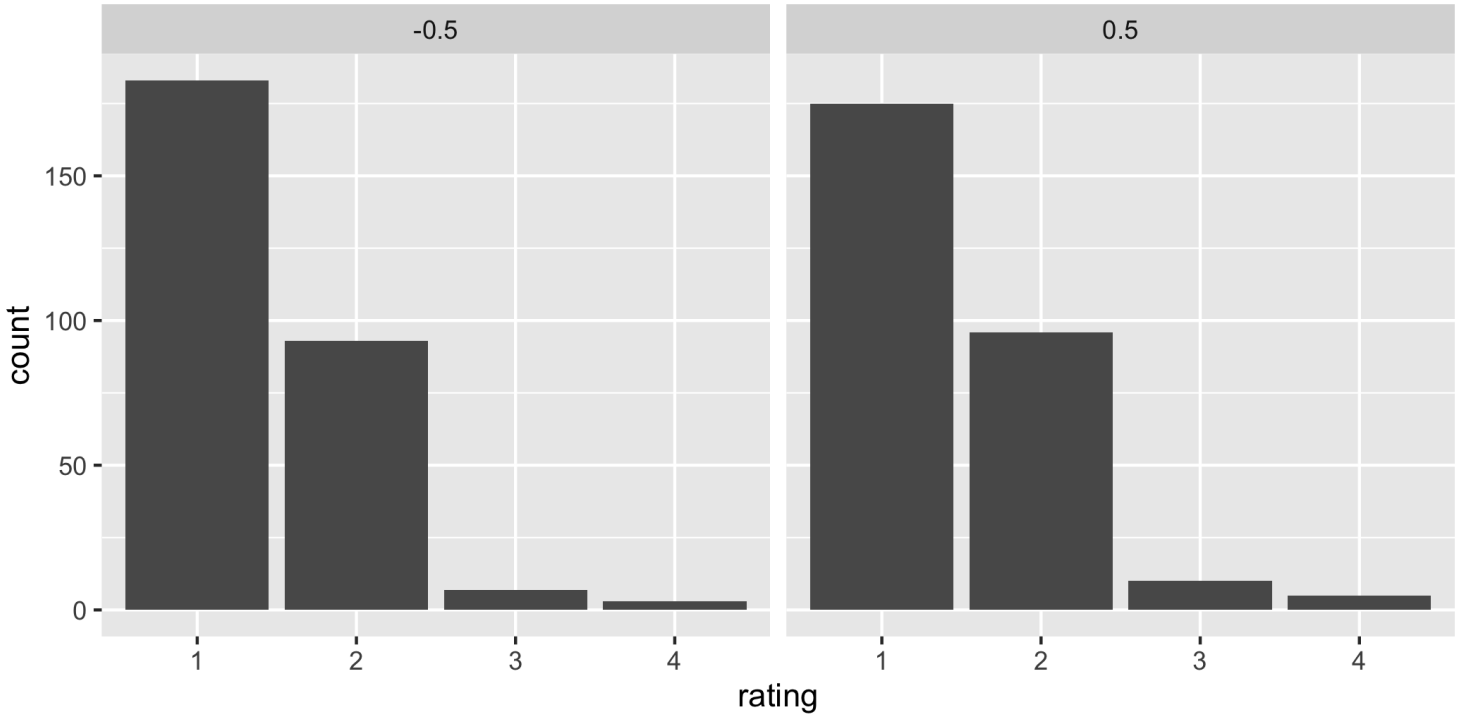
```
ggplot(data=inhaler, aes(x=rating)) +
geom_bar(stat="count")+facet_wrap(~treat)
```



We see equal numbers in each group; it seems that the inhaler insert labeled 0.5 may have been easier to understand.

Clarity of inhaler instructions

```
ggplot(data=inhaler, aes(x=rating)) +
geom_bar(stat="count")+facet_wrap(~period)
```



Period does not seem to have much impact on the ratings.

Model

Let's consider the model

$$\log\left(\frac{\pi_{ijk}}{\pi_{ij1}}\right) = \beta_{0k} + \beta_{1k}t_{ij} + \beta_{2k}p_{ij} + b_{ik}; \quad k = 2, 3, 4;$$
$$b_{ik} \sim N(0, \sigma_k^2).$$

where

- t_{ij} indicates the inhaler insert used by individual i in period j, and
- p_{ij} indicates the corresponding period of measurement.

Implementation in R

```
#Note that these models can take a while to run
#They can also have relatively low ESS
#Default priors:
  #Half t_3 scale 10 on grand intercept,
  #Half t_3, scale 10 on SD,
  #Uniform improper on slopes
m1 <- brm(rating ~ treat + period + (1|subject),
          data=inhaler,
          family=categorical(),
          control=list(adapt_delta=0.99),
          chains=3)
summary(m1)
```

Results

```
##  Family: categorical
##   Links: mu2 = logit; mu3 = logit; mu4 = logit
## Formula: rating ~ treat + period + (1 | subject)
##    Data: inhaler (Number of observations: 572)
## Samples: 3 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 3000
##
## Group-Level Effects:
## ~subject (Number of levels: 286)
##                   Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(mu2_Intercept)     1.26      0.28     0.73     1.83 1.01      502      769
## sd(mu3_Intercept)     2.06      1.18     0.17     4.79 1.01      371      546
## sd(mu4_Intercept)     1.13      0.91     0.04     3.36 1.01      761     1186
##
## Population-Level Effects:
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## mu2_Intercept    -0.37      0.20    -0.76     0.02 1.00     4347     2347
## mu3_Intercept    -4.13      1.43    -7.72    -2.19 1.01      495     1057
## mu4_Intercept    -4.45      1.32    -7.76    -2.77 1.00     1227     1430
## mu2_treat0.5     -1.10      0.22    -1.55    -0.69 1.00     3500     2103
## mu2_period0.5     0.10      0.20    -0.28     0.49 1.00     9175     2227
## mu3_treat0.5     -3.02      1.04    -5.37    -1.31 1.00     1694     1473
## mu3_period0.5     0.30      0.63    -0.98     1.51 1.00     4827     2541
## mu4_treat0.5     -1.67      0.92    -3.75    -0.09 1.00     4412     1580
## mu4_period0.5     0.66      0.82    -0.90     2.48 1.00     5537     1862
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Results

Here we see evidence that when using the inhaler and instructions labeled 0.5, participants are more likely than when using the other inhaler and instructions (labeled -0.5), to select the easy rating than any of the other options.

It's hard to estimate these variance components -- data are sparse for the higher categories.

Ordinal responses

Suppose the categories of our response variable has a natural ordering.

Let's start with data from Example 6.2.2 from Alan Agresti's An Introduction to Categorical Data Analysis, Second Edition to demonstrate this.

This data is from a General Social Survey. Clearly, political ideology has a five-point ordinal scale, ranging from very liberal to very conservative.

		Political Ideology				
		Very Liberal	Slightly Liberal	Moderate	Slightly Conservative	Very Conservative
Female	Democratic	44	47	118	23	32
	Republican	18	28	86	39	48
Male	Democratic	36	34	53	18	23
	Republican	12	18	62	45	51

Cumulative logits

When we have ordinal response with categories $1, 2, \ldots, K$, we still want to estimate

$$\Pr[y_i = 1|\boldsymbol{x}_i] = \pi_{i1}, \ \Pr[y_i = 2|\boldsymbol{x}_i] = \pi_{i2}, \ \ldots, \ \Pr[y_i = K|\boldsymbol{x}_i] = \pi_{iK}.$$

However, we need to use models that can reflect the ordering

$$\Pr[y_i \leq 1|\boldsymbol{x}_i] \leq \Pr[y_i \leq 2|\boldsymbol{x}_i] \leq \ldots \leq \Pr[y_i \leq K|\boldsymbol{x}_i] = 1.$$

Notice that the ordering of probabilities is not for the actual marginal probabilities, but rather the cumulative probabilities.

The multinomial logistic regression does not enforce this.

Instead, we can focus on building models for the cumulative logits, that is, models for

$$\log\left(\frac{\Pr[y_i \leq k|\boldsymbol{x}_i]}{\Pr[y_i > k|\boldsymbol{x}_i]}\right) = \log\left(\frac{\pi_{i1} + \pi_{i2} + \ldots + \pi_{ik}}{\pi_{i(k+1)} + \pi_{i(k+2)} + \ldots + \pi_{iK}}\right), \quad k = 1, \ldots, K-1.$$

Proportional odds model

This leads us to the proportional odds model, written as:

$$\log\left(\frac{\Pr[y_i \leq k|\boldsymbol{x}_i]}{\Pr[y_i > k|\boldsymbol{x}_i]}\right) = \beta_{0k} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \quad k = 1, \ldots, K-1.$$

There is no need to model $\Pr[y_i \leq K]$ since it is necessarily equal to 1.

Notice that this model looks like a binary logistic regression in which we combine the first k categories to form a single category (say 1) and the remaining categories to form a second category (say 0).

Since β_0 is the only parameter indexed by k, the $K-1$ logistic regression curves essentially have the same shapes but different "intercepts".

That is, the effect of the predictors is identical for all $K-1$ cumulative log odds.

This is therefore, a more parsimonious model (both in terms of estimation and interpretation) than the multinomial logistic regression, when it fits the data well.

Proportional odds model

The probabilities we care about are quite easy to extract, since each

$$\Pr[y_i = k|\boldsymbol{x}_i] = \Pr[y_i \leq k|\boldsymbol{x}_i] - \Pr[y_i \leq k-1|\boldsymbol{x}_i], \quad k = 2, \ldots, K,$$

with $\Pr[y_i \leq 1|\boldsymbol{x}_i] = \Pr[y_i = 1|\boldsymbol{x}_i]$.

Let's focus first on a single continuous predictor, that is,

$$\log\left(\frac{\Pr[y_i \leq k|x_i]}{\Pr[y_i > k|x_i]}\right) = \beta_{01} + \beta_1 x_{i1}, \quad k = 1, \ldots, K-1.$$

Here, $\beta_1 > 0$, actually means that a 1 unit increase in x makes the larger values of Y less likely.

This can seem counter-intuitive in many examples, thus, many books and software packages often write

$$\log\left(\frac{\Pr[y_i \leq k|x_i]}{\Pr[y_i > k|x_i]}\right) = \beta_{01} - \beta_1 x_{i1}, \quad k = 1, \ldots, K-1$$

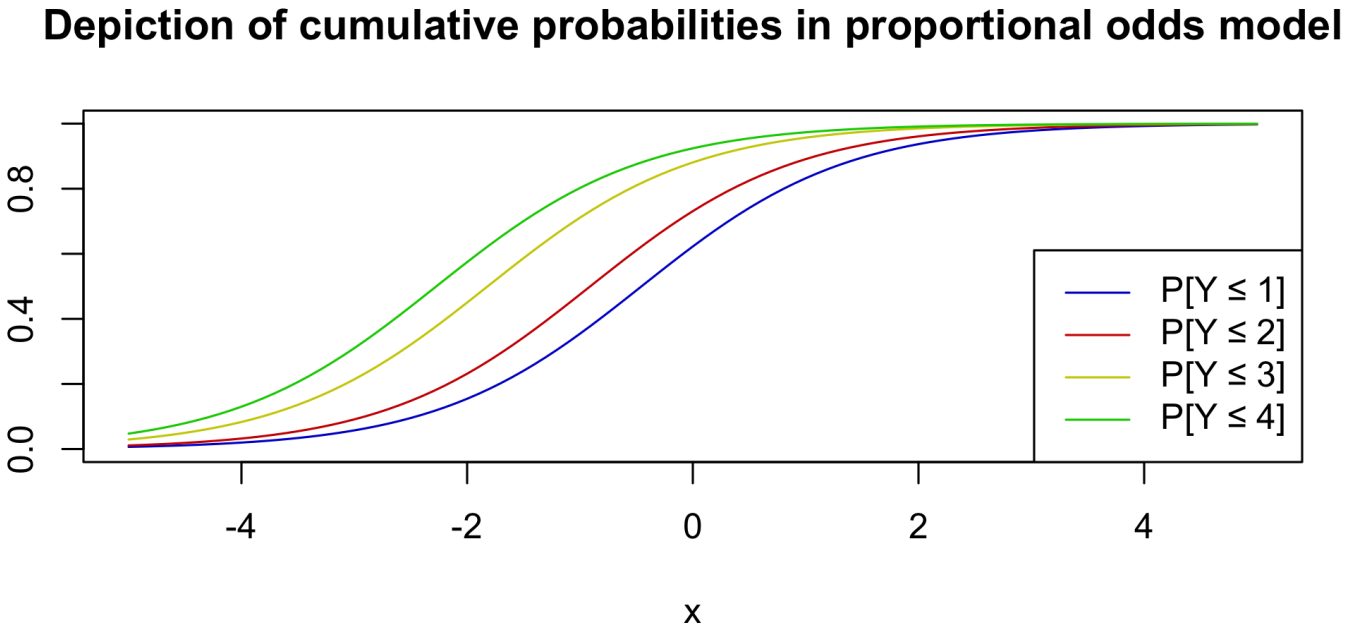
instead. Always check the documentation of your function to ascertain the representation of the model.

Proportional odds model

Suppose we have $K = 5$, $\beta_1 = 1.1$, and $(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}) = (0.5, 1, 2, 2.5)$ in the first representation

$$\log\left(\frac{\Pr[y_i \leq k|x_i]}{\Pr[y_i > k|x_i]}\right) = \beta_{0j} + \beta_1 x_{i1}, \quad k = 1, \ldots, 4,$$

the cumulative probabilities would look like:

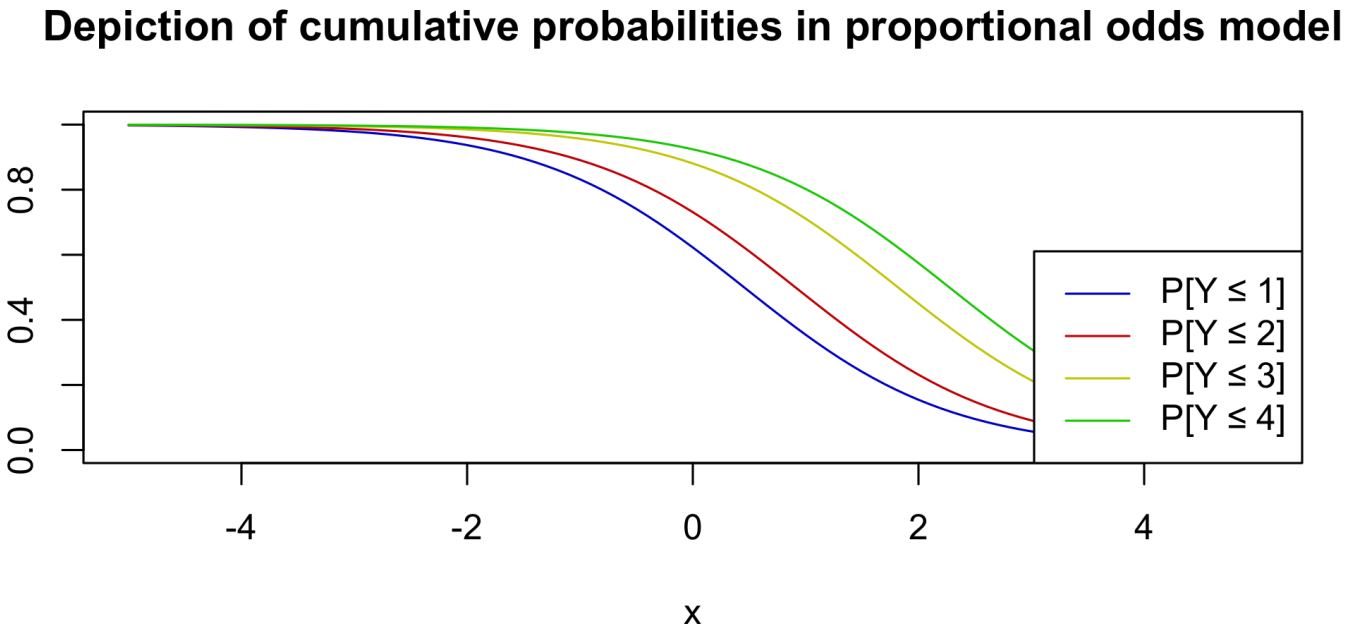


Proportional odds model

But with $K = 5$, and the same values $\beta_1 = 1.1$, and $(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}) = (0.5, 1, 2, 2.5)$ in the second representation

$$\log\left(\frac{\Pr[y_i \leq k|x_i]}{\Pr[y_i > k|x_i]}\right) = \beta_{0j} - \beta_1 x_{i1}, \quad k = 1, \ldots, 4,$$

the cumulative probabilities would look like:



Proportional odds model

Take our example on political ideology for instance. Suppose we fit the model

$$\log\left(\frac{\Pr[\text{ideology}_i \leq k|x_i]}{\Pr[\text{ideology}_i > k|x_i]}\right) = \beta_{0k} - \beta_1 x_{i1}, \quad k = 1, \ldots, 4,$$

where x is an indicator variable for political party, with $x = 1$ for Democrats and $x = 0$ for Republicans.

Then,

- For any k, β_1 is the log-odds of a Democrat, when compared to a Republican, of being more conservative than j compared to being more liberal than j.
- For any k, e^{β_1} is the odds of a Democrat, when compared to a Republican, of being more conservative than j compared to being more liberal than j.

If $\beta_1 > 0$, a Democrat's response is more likely than a Republican's response to be in the conservative direction than in the liberal direction.

Hierarchical extension

Again consider the model

$$\log\left(\frac{\Pr[y_i \leq k|x_i]}{\Pr[y_i > k|x_i]}\right) = \beta_{0k} - \beta_1 x_i, \quad k = 1, \ldots, K-1.$$

Just as before, it is relatively straightforward to consider extensions to this model.

Unlike before however, it makes sense to have one random intercept per person, since we have ordinal responses.

So, we can write

$$\log\left(\frac{\Pr[y_{ij} \leq k|x_{ij}]}{\Pr[y_{ij} > k|x_{ij}]}\right) = \beta_{0k} - \left[\beta_1 x_{ij} + b_i\right]; \quad k = 1, \ldots, K-1;$$

$$b_i \sim N(0, \sigma^2).$$

Back to inhaler data

Recall that the outcome from the inhaler data is actually ordinal.

That is,

- 1 = easy to understand
- 2 = only clear after rereading
- 3 = not very clear
- 4 = confusing.

Thus, it makes sense to also consider a proportional odds model here.

Model

We can then fit the model:

$$\log\left(\frac{\Pr[y_{ij} \leq k|x_{ij}]}{\Pr[y_{ij} > k|x_{ij}]}\right) = \beta_{0k} - \left[\beta_1 t_{ij} + \beta_2 p_{ij} + b_i\right]; \quad k = 1, 2, 3;$$

$$b_i \sim N(0, \sigma^2).$$

where

- t_{ij} indicates the inhaler insert used by individual i in period j, and
- p_{ij} indicates the corresponding period of measurement.

Implementation in R

```
#BRMS follows the convention I mentioned earlier with the -ve slopes
#so need to be careful when interpreting the model
m2 <- brm(rating ~ treat + period + (1|subject),
          data=inhaler,
          family=cumulative(logit),
          control=list(adapt_delta=0.95))
summary(m2)
```

Results

```
##  Family: cumulative
##   Links: mu = logit; disc = identity
## Formula: rating ~ treat + period + (1 | subject)
##    Data: inhaler (Number of observations: 572)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Group-Level Effects:
## ~subject (Number of levels: 286)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.18      0.24     0.71     1.65 1.00      877     1560
##
## Population-Level Effects:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]     0.15      0.18    -0.20     0.52 1.00     8133     3316
## Intercept[2]     3.29      0.31     2.70     3.92 1.00     2502     3094
## Intercept[3]     4.59      0.44     3.76     5.55 1.00     3366     3127
## treat0.5        -1.28      0.21    -1.69    -0.87 1.00     4261     2778
## period0.5        0.20      0.20    -0.19     0.58 1.00     9746     2695
##
## Family Specific Parameters:
##      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## disc     1.00      0.00     1.00     1.00 1.00     4000     4000
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Results

Here we see evidence that when using the inhaler and instructions labeled 0.5, participants are more likely than when using the other inhaler and instructions (labeled -0.5) to select the "easy" rating than any of the other options.

Since $\beta_1 < 0$, that is -1.27, those with the 0.5 inhaler are more likely than to be in the "easy" direction than in the "confusing" direction, those with the -0.5 inhaler.

So we have, with the 0.5 inhaler, participants have 1.27 with CI: $(0.88, 1.69)$ times the odds of picking "easy" versus the other 3 categories.

They also then have 1.27 with CI: $(0.88, 1.69)$ times the odds of picking the first two categories, that is "easy or only clear after rereading", versus the other 2 categories. And so on...

Again, there's not much of a learning effect reflected in the period estimate.

What's next?

Move on to the readings for the next module!