SOC6707 Intermediate Data Analysis

Monica Alexander

Week 10: Interactions, Polytomous outcomes

Notes

- Assignment 3
- Research project analysis
- Plan for remaining weeks
    - today: multinomial
    - next week: miscellaneous research design
    - week 12: presentations

Polytomous outcomes

Polytomous outcomes

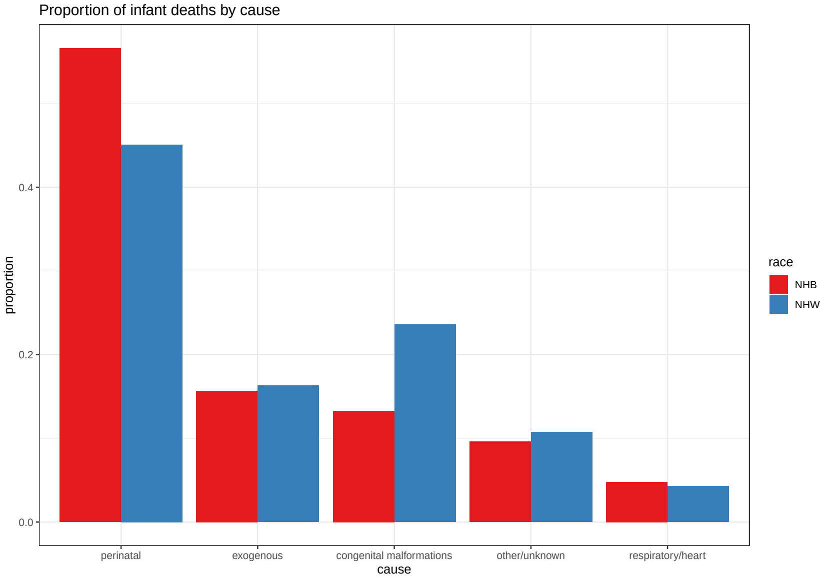
- So far we have only considered continuous and binary response variables, but what if we are interested in modeling a polytomous response variable as a function of continuous and/or categorical explanatory variables?
- A polytomous response variable is a variable that takes on one of $j > 2$ possible values representing membership in one of $j > 2$ different groups or categories. Examples:
  - Self-reported health
  - Voted Liberal, Conservative, NDP, Greens
  - Cause of death
- Polytomous response variables can be ordered or not, and can be modeled in several different ways
- Here I will focus on **multinomial logistic regression**

Multinomial response

- A multinomial variable is a particular type of polytomous variable where the $j > 2$ different groups or categories are not ordered
- Example: cause of infant death in the US. Here's what the dataset looks like:

race	mom_age	gest	preterm	cod_group
NHW	30	27	1	perinatal
NHW	32	36	1	congenital malformations
NHW	25	44	0	perinatal
NHB	29	21	1	perinatal
NHB	23	26	1	perinatal
NHW	34	39	0	congenital malformations

Cause of infant death



Multinomial distribution

- Now Y_i make take one of several discrete values, $1, 2, \ldots, J$.
- Now the probability is

$$\pi_{ij} = Pr(Y_i = j)$$

with

$$\sum_j \pi_{ij} = 1$$

- Note that this is an extension of the binomial distribution (for binary variables), which is the same thing, just with $J = 2$
- As such we can model multinomial outcomes in much the same way, using multinomial logistic regression

Multinomial logistic regression

- Multinomial logistic regression is a model for the conditional probability that a multinomial response variable is equal to j given a set of explanatory variables
- The MLRM can be expressed as

$$\log\left(\frac{P\left(Y_i = j \mid X_{i1}, \ldots, X_{ik}\right)}{P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right)}\right) = \eta_{ji} = \beta_{j0} + \beta_{j1} X_{i1} + \cdots + \beta_{jk} X_{ik} \quad \text{for } j = 1, \ldots,$$

where $\log\left(\frac{P(Y_i = j | X_{i1}, \ldots, X_{ik})}{P(Y_i = 1 | X_{i1}, \ldots, X_{ik})}\right)$ is known as the "log odds of response category 'j' versus response category 1" and β_{jk} are a set of unknown parameters subject to the constraint that $\beta_{1k} = 0$ for all k.

Multinomial logistic regression

$$\log\left(\frac{P\left(Y_i = j \mid X_{i1}, \ldots, X_{ik}\right)}{P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right)}\right) = \eta_{ji} = \beta_{j0}+\beta_{j1}X_{i1}+\cdots+\beta_{jk}X_{ik} \quad \text{for } j = 1, \ldots, J$$

what is category 1?

- Doesn't really matter what it is
- R will by default choose (what)?
- But can change it to be what you want by re-leveling factors

Multinomial logistic regression

Because the logit link function is invertible, we can also express the MLRM as an inverse logit function:

$$
\begin{aligned}
P\left(Y_i = j \mid X_{i1}, \ldots, X_{ik}\right) &= \frac{\exp\left(\eta_{ji}\right)}{\sum_j \exp\left(\eta_{ji}\right)} \\
&= \frac{\exp\left(\beta_{j0} + \beta_{j1} X_{i1} + \cdots + \beta_{jk} X_{ik}\right)}{\sum_j \exp\left(\beta_{j0} + \beta_{j1} X_{i1} + \cdots + \beta_{jk} X_{ik}\right)}
\end{aligned}
$$

Multinomial logistic regression

More specifically, we can express the conditional probabilities as follows:

$$\begin{aligned}
P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right) &= \tfrac{\exp(\eta_{1i})}{\sum_j \exp(\eta_{ji})} = \tfrac{1}{1+\exp(\eta_{2i})+\cdots+\exp(\eta_{Ji})} \\
P\left(Y_i = 2 \mid X_{i1}, \ldots, X_{ik}\right) &= \tfrac{\exp(\eta_{2i})}{\sum_j \exp(\eta_{ji})} = \tfrac{\exp(\eta_{2i})}{1+\exp(\eta_{2i})+\cdots+\exp(\eta_{Ji})} \\
&\vdots \\
P\left(Y_i = J \mid X_{i1}, \ldots, X_{ik}\right) &= \tfrac{\exp(\eta_{Ji})}{\sum_j \exp(\eta_{ji})} = \tfrac{\exp(\eta_{Ji})}{1+\exp(\eta_{2i})+\cdots+\exp(\eta_{Ji})}
\end{aligned}$$

Interpretation

What is the parameter β_{j1} for $j > 1$?

$$
\begin{aligned}
&\log\left(\frac{P\left(Y_i=j|X_{i1}=x_1^*+1,X_{i2}=x_2^*,\ldots,X_{ik}=x_k^*\right)}{P\left(Y_i=1|X_{i1}=x_1^*+1,X_{i2}=x_2^*,\ldots,X_{ik}=x_k^*\right)}\right)-\log\left(\frac{P\left(Y_i=j|X_{i1}=x_1^*,X_{i2}=x_2^*,\ldots,X_{ik}=x_k^*\right)}{P\left(Y_i=1|X_{i1}=x_1^*,X_{i2}=x_2^*,\ldots,X_{ik}=x_k^*\right)}\right)\\
&\quad=\left(\beta_{j0}+\beta_{j1}\left(x_1^*+1\right)+\beta_{j2}x_2^*+\cdots+\beta_{jk}x_k^*\right)-\left(\beta_{j0}+\beta_{j1}x_1^*+\beta_{j2}x_2^*+\cdots+\beta_{jk}x_k^*\right)\\
&\quad=\beta_{j1}
\end{aligned}
$$

β_{j1} is a log odds ratio that gives the change in the log odds that Y_i is equal to j rather than 1 associated with a unit increase in X_{i1}, holding other explanatory variables constant.

Interpretation

What is $\exp(\beta_{j1})$?

$$
\begin{aligned}
\exp\left(\beta_{j1}\right) &= \exp\left(\log\left(\frac{P\left(Y_i = j \mid X_{i1} = x_1^* + 1, \ldots\right)}{P\left(Y_i = 1 \mid X_{i1} = x_1^* + 1, \ldots\right)} \Big/ \frac{P\left(Y_i = j \mid X_{i1} = x_1^*, \ldots, \right)}{P\left(Y_i = 1 \mid X_{i1} = x_1^*, \ldots\right)}\right)\right) \\
&= \frac{P\left(Y_i = j \mid X_{i1} = x_1^* + 1, \ldots\right)}{P\left(Y_i = 1 \mid X_{i1} = x_1^* + 1, \ldots\right)} \Big/ \frac{P\left(Y_i = j \mid X_{i1} = x_1^*, \ldots\right)}{P\left(Y_i = 1 \mid X_{i1} = x_1^*, \ldots\right)}
\end{aligned}
$$

$\exp(\beta_{j1})$ is the odds ratio that gives the multiplicative change in the odds that Y_i is equal to j rather than 1 associated with a unit increase in X_{i1}, holding other explanatory variables constant.

Comparing other response categories

The preceding calculations concerned the contrast between response category j and the baseline category 1, but they are easily extended to contrasts between any two categories j and j'

Specifically, the log odds ratio that Y_i is equal to j rather than j' associated with a unit increase in X_{ik}, holding other variables constant, is

$$\log\left(\frac{P\left(Y_i = j \mid X_{i1} = x_1^* + 1 \ldots\right)}{P\left(Y_i = j' \mid X_{i1} = x_1^* + 1, \ldots\right)} \Big/ \frac{P\left(Y_i = j \mid X_{i1} = x_1^*, \ldots\right)}{P\left(Y_i = j' \mid X_{i1} = x_1^*, \ldots\right)}\right) = \beta_{jk} - \beta_{j'k}$$

and the corresponding odds ratio is

$$\frac{P\left(Y_i = j \mid X_{i1} = x_1^* + 1, \ldots\right)}{P\left(Y_i = j' \mid X_{i1} = x_1^* + 1, \ldots\right)} \Big/ \frac{P\left(Y_i = j \mid X_{i1} = x_1^*, \ldots\right)}{P\left(Y_i = j' \mid X_{i1} = x_1^*, \ldots\right)} = \exp\left(\beta_{jk} - \beta_{j'k}\right)$$

General interpretations: take-away

Lots of symbols, but:

- interpretation of coefficients is direct extension of logistic
    - instead of “odds of yes versus no” it’s “odds of thing outcome happening versus another outcome happening”
- so e.g. instead of “odds of dying versus not” it’s “odds of dying from exogenous causes versus perinatal causes”

Maybe the trickiest bit is to get everything into the right format to run the regression

- We had data in long format, but we need summaries in wide format

Example

Get data in wide format. Firstly, get the counts by covariate groups:

```
infant_counts <- infant %>%
  group_by(race, mom_age, gest, preterm, cod_group) %>%
  tally(name = "deaths")
infant_counts
```

```
## # A tibble: 4,113 x 6
## # Groups:   race, mom_age, gest, preterm [1,602]
##    race  mom_age  gest preterm cod_group     deaths
##    <chr>   <dbl> <dbl>   <dbl> <chr>          <int>
##  1 NHB        14    19       1 perinatal          1
##  2 NHB        14    21       1 perinatal          1
##  3 NHB        14    22       1 perinatal          1
##  4 NHB        14    23       1 perinatal          1
##  5 NHB        14    24       1 exogenous          1
##  6 NHB        14    24       1 other/unknown      1
##  7 NHB        14    24       1 perinatal          3
##  8 NHB        14    25       1 perinatal          1
##  9 NHB        14    27       1 other/unknown      1
## 10 NHB        14    27       1 perinatal          2
## # ... with 4,103 more rows
```

Example

Now get in wide format

```
infant_wide <- infant_counts %>%
  pivot_wider(names_from = cod_group, values_from = deaths) %>%
  mutate_all(.funs = funs(ifelse(is.na(.), 0, .)))
head(infant_wide)
```

```
## # A tibble: 6 x 9
## # Groups:   race, mom_age, gest, preterm [6]
##   race  mom_age  gest preterm perinatal exogenous ‘other/unknown‘
##   <chr>   <dbl> <dbl>   <dbl>     <dbl>     <dbl>           <dbl>
## 1 NHB        14    19       1         1         0               0
## 2 NHB        14    21       1         1         0               0
## 3 NHB        14    22       1         1         0               0
## 4 NHB        14    23       1         1         0               0
## 5 NHB        14    24       1         3         1               1
## 6 NHB        14    25       1         1         0               0
## # ... with 2 more variables: congenital malformations <dbl>,
## #   respiratory/heart <dbl>
```

Example

Create outcome Y which is a vector of cause-specific deaths

```r
infant_wide$Y <- as.matrix(infant_wide[,c("perinatal",
                                          "exogenous",
                                          "congenital malformations",
                                          "respiratory/heart", "other/unknown")])
head(infant_wide$Y)
```

```
##      perinatal exogenous congenital malformations respiratory/heart
## [1,]         1         0                        0                 0
## [2,]         1         0                        0                 0
## [3,]         1         0                        0                 0
## [4,]         1         0                        0                 0
## [5,]         3         1                        0                 0
## [6,]         1         0                        0                 0
##      other/unknown
## [1,]             0
## [2,]             0
## [3,]             0
## [4,]             0
## [5,]             1
## [6,]             0
```

Example

```
library(nnet)
mod_mn <- multinom(Y ~ race+ mom_age+ preterm, data = infant_wide)
```

```
## # weights:  25 (16 variable)
## initial  value 27399.071021
## iter  10 value 20149.661320
## iter  20 value 19437.349750
## final  value 19436.462463
## converged
```

```
summary(mod_mn)
```

```
## Call:
## multinom(formula = Y ~ race + mom_age + preterm, data = infant_wide)
##
## Coefficients:
##                          (Intercept)      raceNHW     mom_age   preterm
## exogenous                 2.56320808  0.088345261 -0.05692035 -3.429460
## congenital malformations -0.01647076  0.621524245  0.01916732 -2.423940
## respiratory/heart        -0.15823646 -0.004845986 -0.01780013 -2.251658
## other/unknown             1.10771251  0.145290756 -0.02245255 -3.137589
##
## Std. Errors:
##                          (Intercept)    raceNHW     mom_age    preterm
## exogenous                  0.1235975 0.05354744 0.004365804 0.06000498
## congenital malformations   0.1093744 0.04840430 0.003501902 0.05449309
## respiratory/heart          0.1811151 0.07928810 0.006287607 0.08451183
## other/unknown              0.1361523 0.06022037 0.004717569 0.06546394
##
## Residual Deviance: 38872.92
## AIC: 38904.92
```

Some interpretations

```
coef(mod_mn)
```

```
##                          (Intercept)      raceNHW     mom_age   preterm
## exogenous                 2.56320808  0.088345261 -0.05692035 -3.429460
## congenital malformations -0.01647076  0.621524245  0.01916732 -2.423940
## respiratory/heart        -0.15823646 -0.004845986 -0.01780013 -2.251658
## other/unknown             1.10771251  0.145290756 -0.02245255 -3.137589
```

```
exp(coef(mod_mn))
```

```
##                          (Intercept)   raceNHW   mom_age    preterm
## exogenous                 12.9773831 1.0923652 0.9446693 0.03240443
## congenital malformations   0.9836641 1.8617637 1.0193522 0.08857195
## respiratory/heart          0.8536479 0.9951657 0.9823574 0.10522463
## other/unknown              3.0274253 1.1563757 0.9777976 0.04338727
```

- The odds of exogenous causes compared to perinatal causes for NHW babies is 9% more than NHB babies, holding everything else constant
- The odds of respiratory/heart causes compared to perinatal causes for preterm babies is 90% less than for non-preterm babies, holding everything else constant
- The odds of respiratory/heart causes compared to congenital malformations for preterm babies is $\exp(-2.25 + 2.42) = 1.18$ times (or 18% more) than for non-preterm babies, holding everything else constant

Predicted probabilities

- Now let's convert some of these coefficients into predicted probabilities
- We age, race, and preterm as covariates
- Easiest to pick an age (hold the continuous variable constant at a particular value) and then get predicted probabilities for all other groups
- Create a new tibble with all the groups we want predicted probabilities for:

```
predict_df <- tibble(race = rep(c("NHW", "NHB"), each = 2),
       mom_age = 30,
       preterm = rep(c(0,1),2))
predict_df
```

```
## # A tibble: 4 x 3
##   race  mom_age preterm
##   <chr>   <dbl>   <dbl>
## 1 NHW        30       0
## 2 NHW        30       1
## 3 NHB        30       0
## 4 NHB        30       1
```

Predicted probabilities

Now get the predictions

```
preds <- as_tibble(predict(mod_mn, newdata = predict_df, type = 'probs'))
preds
```

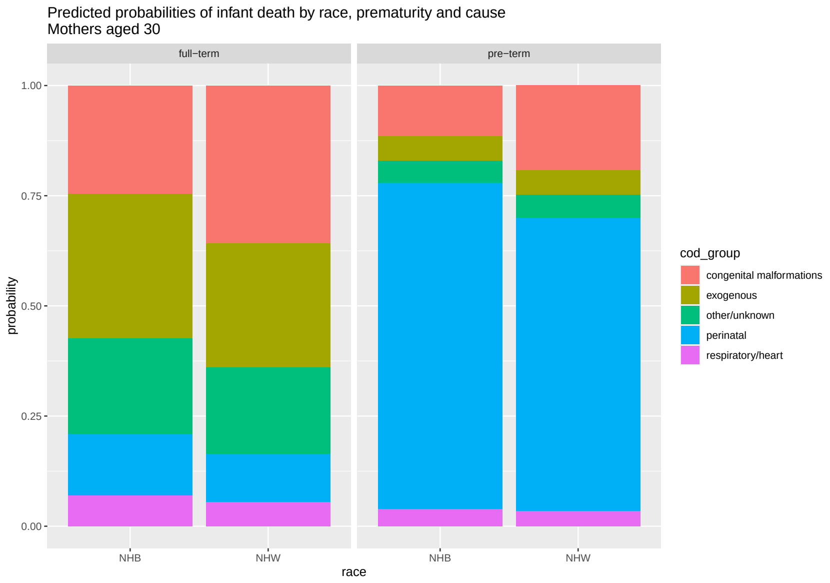
```
## # A tibble: 4 x 5
##   perinatal exogenous `congenital malformati~ `respiratory/hear~ `other/unknown`
##       <dbl>     <dbl>                   <dbl>              <dbl>           <dbl>
## 1     0.110    0.282                    0.357             0.0547          0.196
## 2     0.666    0.0555                   0.192             0.0349          0.0516
## 3     0.140    0.329                    0.245             0.0700          0.216
## 4     0.740    0.0564                   0.115             0.0390          0.0496
```

And join these back onto the tibble:

```
preds <- bind_cols(predict_df, preds)
preds
```

```
## # A tibble: 4 x 8
##   race  mom_age preterm perinatal exogenous `congenital malfor~ `respiratory/he~
##   <chr>   <dbl>   <dbl>     <dbl>     <dbl>               <dbl>            <dbl>
## 1 NHW        30       0     0.110    0.282                0.357           0.0547
## 2 NHW        30       1     0.666    0.0555               0.192           0.0349
## 3 NHB        30       0     0.140    0.329                0.245           0.0700
## 4 NHB        30       1     0.740    0.0564               0.115           0.0390
## # ... with 1 more variable: other/unknown <dbl>
```

Predicted probabilities



Summary

- Multinomial logistic regression is a natural extension of binomial logistic regression
- Useful when you have categorical outcomes with more than 2 categories

A few words on generalized linear models

- So far we've seen linear regression (continuous), logistic regression (binary), and multinomial regression (categorical)
- Notice that all models are of the form

$$g(E(Y_i)) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

where $g(.)$ is some function.

- For linear regression $g(.)$ is the identity
- For logistic regression $g(.)$ is the logit function
- For multinomial regression $g(.)$ is the log of the ratios of probabilities

Generalized linear models

- These are all special cases of generalized linear models (GLM)
- With the appropriate link function $g(.)$, a whole range of variables can be modeled in a linear framework
- We've looked at outcome variables with Normal, Binomial and Multinomial distributions
- But variables from any exponential distribution (a special family of distributions) can be modeled using GLMs
- Other common examples include Poisson, Gamma, and Negative Binomial regression