SOC6707 Intermediate Data Analysis

Monica Alexander

Week 7: Interactions, Polytomous outcomes

Overview

- Correction
- Interaction terms
- Polytomous outcomes

Notes

- No Assignment 2
- But EDA due next week
- Data exploration and write up based on your research question and data set

Outlook:

- Bayesian probability and inference
- Multilevel models

Correction from lab

```
##
## Call:
## glm(formula = low_income ~ age_group + educ_cat, family = "binomial",
##     data = gss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9563  -0.8329  -0.6442   1.0927   2.1556
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -0.82914    0.05256 -15.775  < 2e-16 ***
## age_group10                         4.51050    0.29590  15.244  < 2e-16 ***
## age_group20                         1.03170    0.06369  16.199  < 2e-16 ***
## age_group40                        -0.19260    0.06533  -2.948   0.0032 **
## age_group50                         0.05101    0.05886   0.867   0.3862
## age_group60                         0.34767    0.05642   6.162 7.19e-10 ***
## age_group70                         0.56057    0.06178   9.074  < 2e-16 ***
## age_group80                         0.35590    0.07851   4.533 5.81e-06 ***
## educ_catbachelor                   -0.79571    0.05187 -15.340  < 2e-16 ***
## educ_catless than high school       0.67568    0.05152  13.116  < 2e-16 ***
## educ_catpostgraduate               -1.19840    0.07440 -16.107  < 2e-16 ***
## educ_catsome university            -0.90383    0.10005  -9.034  < 2e-16 ***
## educ_cattrade or other certificate -0.39902    0.04270  -9.345  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25629  on 20260  degrees of freedom
## Residual deviance: 22391  on 20248  degrees of freedom
##   (341 observations deleted due to missingness)
## AIC: 22417
##
## Number of Fisher Scoring iterations: 6
```

Correction

- I incorrectly said the reference category was related to both education and age
- Interpretation for education: “odds of low income is XX compared to high school, holding age constant”
- Interpretation for age: “odds of low income is XX compared to 30-40 year olds, holding education constant”
- I was thinking we had an interaction term. What is an interaction term?

Interaction terms

Effect moderation

- Effect moderation refers to the situation where the partial effect of one explanatory variable differs or changes across levels of another explanatory variable
  - e.g. the association between income and age may vary by education level
- All of the models we have considered thus far constrain the partial effects of the explanatory variables to be invariant, but this may not be appropriate
- If a model constrains partial effects to be invariant when in fact they are not, the estimator is biased for the CEF (our estimates are wrong)

We can accommodate effect moderation through the use of **interaction terms**

Interaction terms

Example of an MLR model with an interaction term:

$$
\begin{aligned}
Y_i &= E\left(Y_i \mid X_{i1}, X_{i2}\right) + \varepsilon_i \\
&= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}
\end{aligned}
$$

- How should we interpret the parameters in an MLR model with interaction terms?
- First, let's take a look at how $E\left(Y_i \mid X_{i1}, X_{i2}\right)$ changes with a unit increase in X_{i1}

Interaction terms

$$E\left(Y_i \mid X_{i1}, X_{i2}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$$

In this model, the change in the expected value of Y_i associated with a unit increase in X_{i1} is given by

$$E\left(Y_i \mid X_{i1} = x_1 + 1, X_{i2} = x_2\right) - E\left(Y_i \mid X_{i1} = x_1, X_{i2} = x_2\right) = \beta_1 + \beta_3 x_2$$

- The partial effect of X_{i1} now depends on the value to which we set the other explanatory variable, X_{i2}
- Note that when $X_{i2} = 0$, this expression simplifies to β_1, or in other words, β_1 is the change in the expected value of Y_i associated with a unit increase in X_{i1} specifically when $X_{i2} = 0$

Interaction terms

Next, let's take a look at how the partial effect of $X_{i1}, \beta_1 + \beta_3 x_2$, changes with a unit increase in X_{i2}

The change in the partial effect of X_{i1} associated with a unit increase in X_{i2} is given by

$$
\begin{aligned}
&[E(Y_i \mid X_{i1} = x_1 + 1, X_{i2} = x_2 + 1) - E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2 + 1)] \\
&- [E(Y_i \mid X_{i1} = x_1 + 1, X_{i2} = x_2) - E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2)] = \beta_3
\end{aligned}
$$

In words, β_3 represents the amount by which the partial effect of X_{i1} differs across levels of the other explanatory variable, X_{i2}

Interaction terms

- The previous slides may take a little getting used to
- In reality, one of our explanatory variables (say X_{i2}) is a binary variable (so either 0 or 1)
- This simplifies the interpretation of the interaction term

Example

- What is the association between TFR, life expectancy and region?
- Does the association between TFR and life expectancy differ based on whether country is in Developed Regions or not?

Example in R

```r
country_ind_2017 <- country_ind %>%
  filter(year==2017) %>%
  mutate(dev_region = ifelse(region=="Developed regions", "yes", "no"))

summary(lm(tfr ~ life_expectancy + dev_region + life_expectancy*dev_region, data = country_ind_2017))
```

```
##
## Call:
## lm(formula = tfr ~ life_expectancy + dev_region + life_expectancy *
##     dev_region, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23326 -0.29618 -0.02426  0.28744  2.54832
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 13.52646    0.52158  25.933  < 2e-16 ***
## life_expectancy             -0.14454    0.00722 -20.019  < 2e-16 ***
## dev_regionyes              -12.95159    2.91594  -4.442 1.59e-05 ***
## life_expectancy:dev_regionyes  0.15711    0.03557   4.417 1.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6164 on 172 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7745
## F-statistic: 201.4 on 3 and 172 DF,  p-value: < 2.2e-16
```

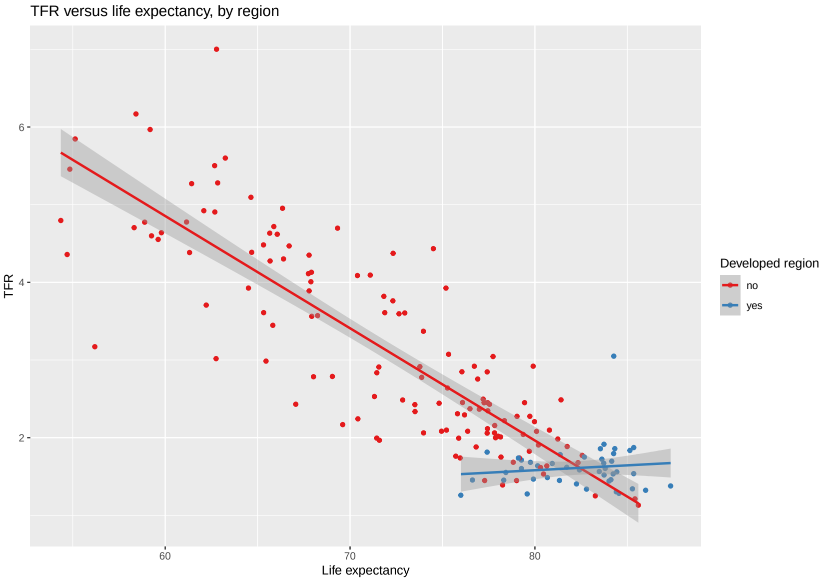
Example

$$Y_i = 13.5 - 0.14X_1 - 13.0X_2 + 0.16X_1X_2$$

Some interpretations

- for non-developed regions, 1 year increase in life expectancy associated with 0.14 decrease in TFR
- for developed regions, a 1 year increase in life expectancy associated with a 0.02 increase in TFR

Visualizing interactions



Back to example from lab

Let's run a simplified version

```r
gss <- gss %>%
  mutate(age_over_30 = ifelse(age>30, "Yes", "No"))
mod_6 <- glm(low_income~ age_over_30 + has_bachelor_or_higher, data = gss, family = "binomial")
summary(mod_6)
```

```
##
## Call:
## glm(formula = low_income ~ age_over_30 + has_bachelor_or_higher,
##     family = "binomial", data = gss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4389  -0.9095  -0.5851   1.3570   1.9232
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                0.59656    0.04095   14.57   <2e-16 ***
## age_over_30Yes            -1.26539    0.04344  -29.13   <2e-16 ***
## has_bachelor_or_higherYes -1.00944    0.03933  -25.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25629  on 20260  degrees of freedom
## Residual deviance: 23973  on 20258  degrees of freedom
##   (341 observations deleted due to missingness)
## AIC: 23979
##
## Number of Fisher Scoring iterations: 4
```

Interpretation

```r
exp(coef(mod_6))
```

```
##               (Intercept)            age_over_30Yes has_bachelor_or_higherYes
##                 1.8158593                 0.2821281                 0.3644247
```

- Odds of low income for 30+ year olds is 72% less than <30, holding education constant
- Odds of low income for those with at least a bachelor is 64% less than those without, holding age constant

Now add an interaction

```r
mod_7 <- glm(low_income~ age_over_30 + has_bachelor_or_higher+ age_over_30:has_bachelor_or_higher, data =
summary(mod_7)
```

```
##
## Call:
## glm(formula = low_income ~ age_over_30 + has_bachelor_or_higher +
##     age_over_30:has_bachelor_or_higher, family = "binomial",
##     data = gss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4527  -0.9076  -0.5911   1.4038   1.9136
##
## Coefficients:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                0.62716    0.04566  13.737   <2e-16
## age_over_30Yes                            -1.30124    0.04940 -26.339   <2e-16
## has_bachelor_or_higherYes                 -1.14520    0.09646 -11.872   <2e-16
## age_over_30Yes:has_bachelor_or_higherYes   0.16308    0.10557   1.545    0.122
##
## (Intercept)                               ***
## age_over_30Yes                            ***
## has_bachelor_or_higherYes                 ***
## age_over_30Yes:has_bachelor_or_higherYes
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25629  on 20260  degrees of freedom
## Residual deviance: 23970  on 20257  degrees of freedom
##   (341 observations deleted due to missingness)
## AIC: 23978
##
## Number of Fisher Scoring iterations: 4
```

Interpretation

```r
coef(mod_7)
```

```
##                              (Intercept)
##                                0.6271583
##                           age_over_30Yes
##                               -1.3012437
##                  has_bachelor_or_higherYes
##                               -1.1451977
## age_over_30Yes:has_bachelor_or_higherYes
##                                0.1630758
```

- For people without a bachelor, odds of low income for 30+ is (1- exp(-1.3))x 100 % = 73% less than for <30 year olds
- For people with a bachelor, odds of low income for 30+ is (1- exp(-1.3+0.16))x 100 % = 70% less than for <30 year olds
- For <30 year olds, odds of low income for bachelor is (1- exp(-1.15))x 100 % = 69% less than for <bachelor
- For 30+ year olds, odds of low income for bachelor is (1- exp(-1.15 + 0.16))x 100 % = 63% less than for <bachelor
- ... but interaction term isn't significant

Polytomous outcomes

Polytomous outcomes

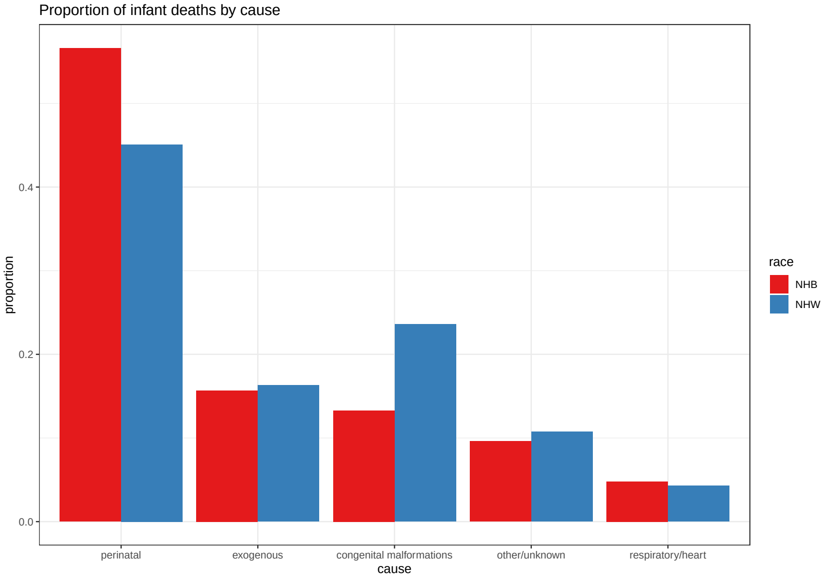
- So far we have only considered continuous and binary response variables, but what if we are interested in modeling a polytomous response variable as a function of continuous and/or categorical explanatory variables?
- A polytomous response variable is a variable that takes on one of $j > 2$ possible 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**, which is for unordered outcomes

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

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

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

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:

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

Interpretation

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

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

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

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

Example

First step: get data in wide format

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

```
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

```r
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
```

```r
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

```r
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
```

```r
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

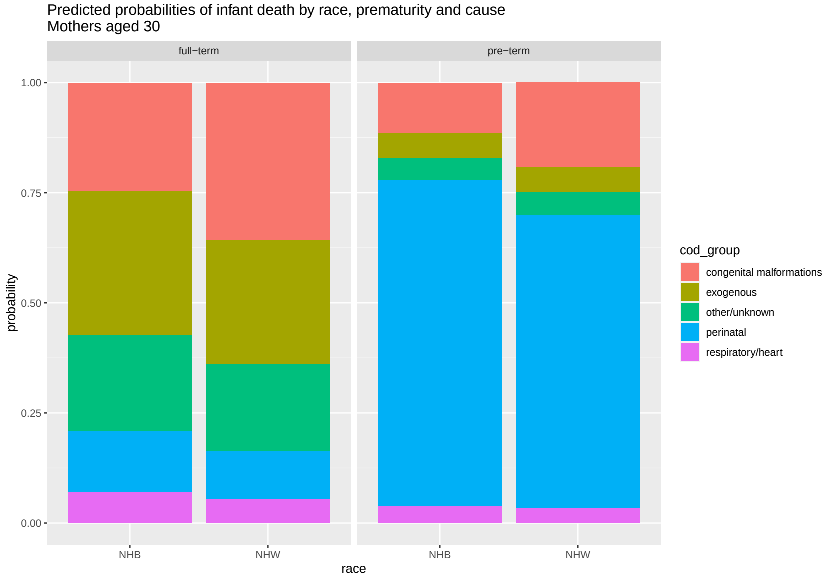
Predicted probabilities

For mothers of age 30

```
predict_df <- tibble(race = rep(c("NHW", "NHB"), each = 2),
       mom_age = 30,
       preterm = rep(c(0,1),2))
preds <- bind_cols(predict_df, as_tibble(predict(mod_mn, newdata = predict_df, type = 'probs')))
preds
```

```
## # A tibble: 4 x 8
##   race  mom_age preterm perinatal exogenous `congenital mal~ `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 (what we saw in week 5)
- Useful when you have categorical outcomes with more than 2 categories
- If the categories are ordered, it's also possible to do **ordered logistic regression**
- Not talked about today, but happy to chat offline if useful for research projects

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