Multinomial Logistic Regression

The Basics

Prof. Maria Tackett

04.06.20

STA 210

[Click for PDF of slides]

Generalized Linear Models (GLM)

- In practice, there are many different types of response variables including:
    - **Binary**: Win or Lose
    - **Nominal**: Democrat, Republican or Third Party candidate
    - **Ordered**: Movie rating (1 - 5 stars)
    - and others...
- These are all examples of **generalized linear models**, a broader class of models that generalize the multiple linear regression model
- See *Generalized Linear Models: A Unifying Theory* for more details about GLMs

Binary Response (Logistic)

- Given $P(y_i = 1|x_i) = \hat{\pi}_i \quad \text{and} \quad P(y_i = 0|x_i) = 1 - \hat{\pi}_i$

$$\log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- We can calculate $\hat{\pi}_i$ by solving the logit equation:

$$\hat{\pi}_i = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\}}$$

Binary Response (Logistic)

- Suppose we consider $y = 0$ the baseline category such that

$$P(y_i = 1|x_i) = \pi_{i1} \;\; \text{and} \;\; P(y_i = 0|x_i) = \pi_{i0}$$

- Then the logistic regression model is

$$\log\left(\frac{\hat{\pi}_{i1}}{1 - \hat{\pi}_{i1}}\right) = \log\left(\frac{\hat{\pi}_{i1}}{\hat{\pi}_{i0}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Slope, $\hat{\beta}_1$: When x increases by one unit, the predicted odds of $y = 1$ versus the baseline $y = 0$ are multiply by a factor of $\exp\{\hat{\beta}_1\}$

- Intercept, $\hat{\beta}_0$: When $x = 0$, the predicted odds of $y = 1$ versus the baseline $y = 0$ are $\exp\{\hat{\beta}_0\}$

STA 210

Multinomial response variable

- Suppose the response variable y is categorical and can take values $1, 2, \dots, K$ such that $(K > 2)$
- **Multinomial Distribution:**

$$P(y = 1) = \pi_1, P(y = 2) = \pi_2, \dots, P(y = K) = \pi_K$$

such that $\sum\limits_{k=1}^{K} \pi_k = 1$

Multinomial Logistic Regression

- If we have an explanatory variable x, then we want to fit a model such that $P(y = k) = \pi_k$ is a function of x

- Choose a baseline category. Let's choose $y = 1$. Then,

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

- In the multinomial logistic model, we have a separate equation for each category of the response relative to the baseline category
  - If the response has K possible categories, there will be $K - 1$ equations as part of the multinomial logistic model

STA 210

Multinomial Logistic Regression

- Suppose we have a response variable y that can take three possible outcomes that are coded as "A", "B", "C"

- Let "A" be the baseline category. Then

$$
\log\left(\frac{\pi_{iB}}{\pi_{iA}}\right) = \beta_{0B} + \beta_{1B}x_i
$$

$$
\log\left(\frac{\pi_{iC}}{\pi_{iA}}\right) = \beta_{0C} + \beta_{1C}x_i
$$

STA 210

NHANES Data

- National Health and Nutrition Examination Survey is conducted by the National Center for Health Statistics (NCHS)
- The goal is to *"assess the health and nutritional status of adults and children in the United States"*
- This survey includes an interview and a physical examination

STA 210

NHANES Data

- We will use the data from the `NHANES` R package
- Contains 75 variables for the 2009 - 2010 and 2011 - 2012 sample years
- The data in this package is modified for educational purposes and should **not** be used for research
- Original data can be obtained from the NCHS website for research purposes
- Type `?NHANES` in console to see list of variables and definitions

STA 210

NHANES: Health Rating vs. Age & Physical Activity

- **Question**: Can we use a person's age and whether they do regular physical activity to predict their self-reported health rating?
- We will analyze the following variables:
  - **`HealthGen`:** Self-reported rating of participant's health in general. Excellent, Vgood, Good, Fair, or Poor.
  - **`Age`:** Age at time of screening (in years). Participants 80 or older were recorded as 80.
  - **`PhysActive`:** Participant does moderate to vigorous-intensity sports, fitness or recreational activities

STA 210

The data

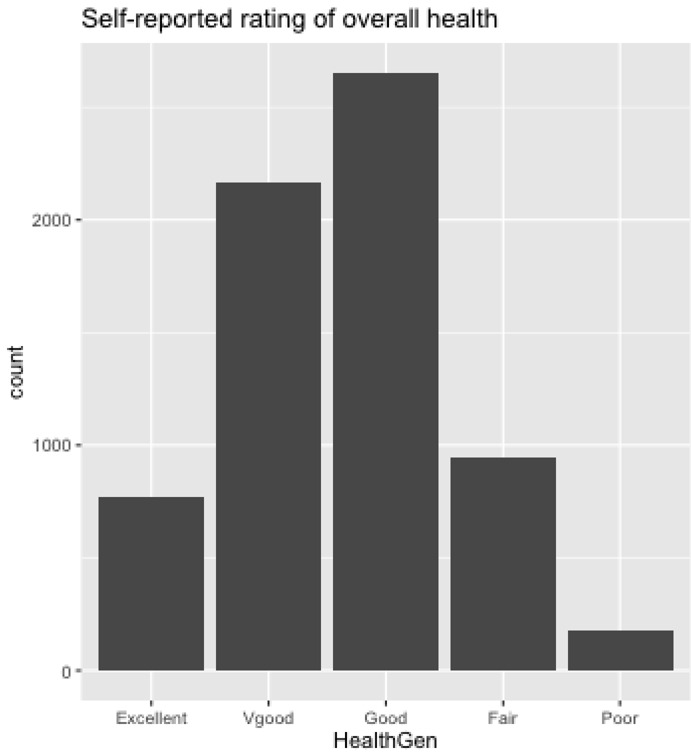
```
library(NHANES)

nhanes_adult <- NHANES %>%
  filter(Age >= 18) %>%
  select(HealthGen, Age, PhysActive) %>%
  drop_na() %>%
  mutate(obs_num = 1:n())
```

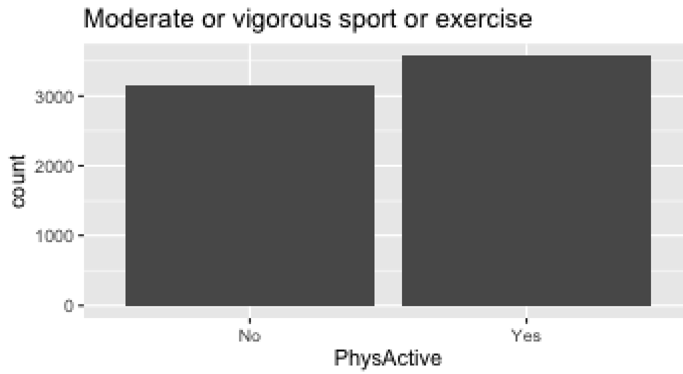
```
glimpse(nhanes_adult)
```

```
## Observations: 6,710
## Variables: 4
## $ HealthGen  <fct> Good, Good, Good, Good, Vgood, Vgood, Vgood, Vgood, V
## $ Age        <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 50, 33, 60, 5
## $ PhysActive <fct> No, No, No, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No
## $ obs_num    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
```

STA 210

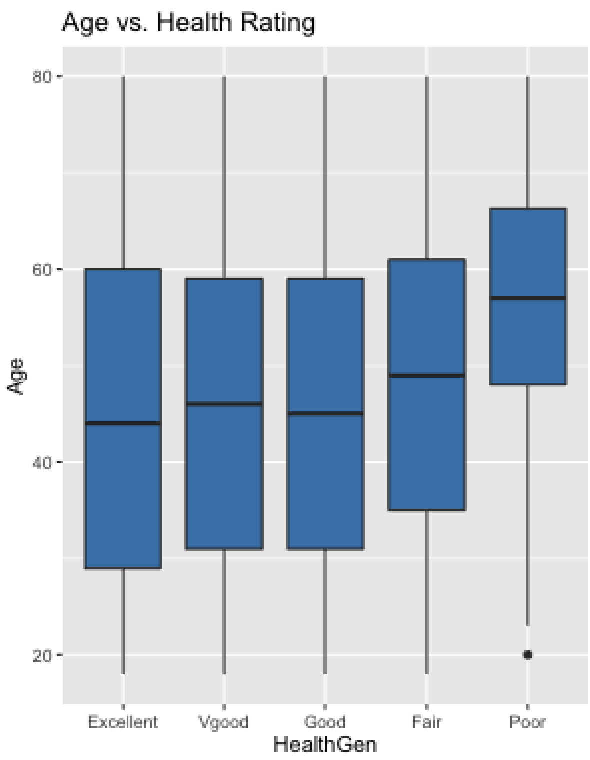
Exploratory data analysis

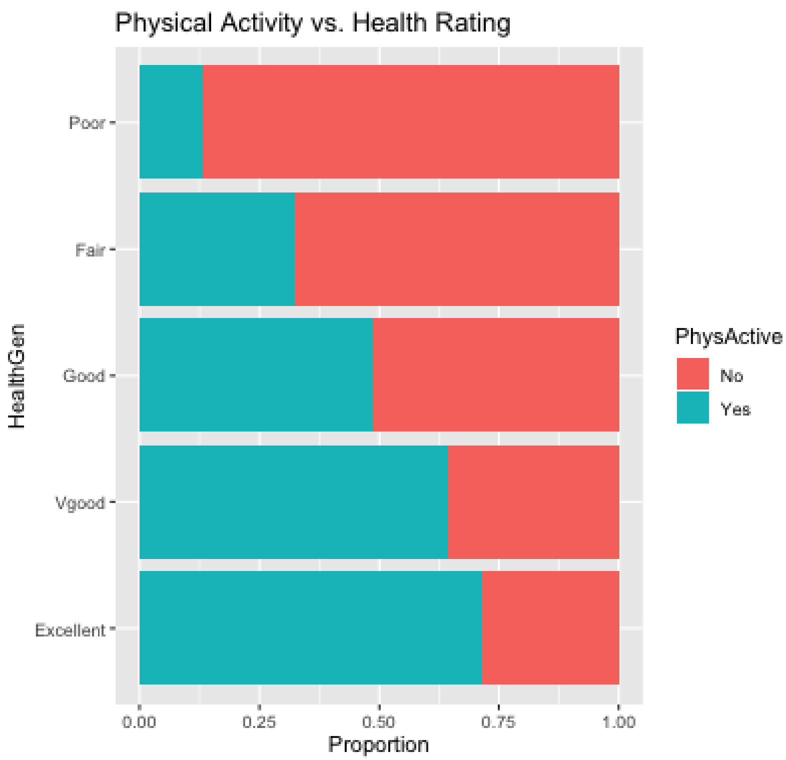




STA 210

Exploratory data analysis





STA 210

Model in R

- Use the **`multinom()`** function in the `nnet` package

```
library(nnet)
health_m <- multinom(HealthGen ~ Age + PhysActive,
                     data = nhanes_adult)
```

- Put `results = "hide"` in the code chunk header to suppress convergence output

STA 210

HealthGen vs. Age and PhysActive

```
tidy(health_m, conf.int = TRUE, exponentiate = FALSE) %>%
  kable(digits = 3, format = "markdown")
```

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Vgood	(Intercept)	1.205	0.145	8.325	0.000	0.922	1.489
Vgood	Age	0.001	0.002	0.369	0.712	-0.004	0.006
Vgood	PhysActiveYes	-0.321	0.093	-3.454	0.001	-0.503	-0.139
Good	(Intercept)	1.948	0.141	13.844	0.000	1.672	2.223
Good	Age	-0.002	0.002	-0.977	0.329	-0.007	0.002
Good	PhysActiveYes	-1.001	0.090	-11.120	0.000	-1.178	-0.825
Fair	(Intercept)	0.915	0.164	5.566	0.000	0.592	1.237
Fair	Age	0.003	0.003	1.058	0.290	-0.003	0.009
Fair	PhysActiveYes	-1.645	0.107	-15.319	0.000	-1.856	-1.435
Poor	(Intercept)	-1.521	0.290	-5.238	0.000	-2.090	-0.952
Poor	Age	0.022	0.005	4.522	0.000	0.013	0.032
Poor	PhysActiveYes	-2.656	0.236	-11.275	0.000	-3.117	-2.194

STA 210

Interpreting coefficients

1. What is the baseline category for the model?

2. Write the model for the odds that a person rates themselves as having "Fair" health versus the baseline category.

3. Interpret the coefficient of `Age` in terms of the odds that a person rates themselves as having "Poor" health versus baseline category.

4. Interpret the coefficient of `PhysActiveYes` in terms of the odds that a person rates themselves as having "Very Good" health versus baseline category.

Hypothesis test for β_{jk}

Let $y = 1$ be the baseline category for the model.

$$\log\left(\frac{\hat{\pi}_{ik}}{\hat{\pi}_{i1}}\right) = \hat{\beta}_{0k} + \hat{\beta}_{1k}x_{i1} + \dots + \hat{\beta}_{pk}x_{ip}$$

The test of significance for the coefficient β_{jk} is

Hypotheses: $H_0: \beta_{jk} = 0 \text{ vs } H_a: \beta_{jk} \neq 0$

Test Statistic:

$$z = \frac{\hat{\beta}_{jk} - 0}{SE(\hat{\beta}_{jk})}$$

P-value: $P(|Z| > |z|)$,

where $Z \sim N(0,1)$, the Standard Normal distribution

Confidence interval for β_{jk}

- We can calculate the **C\% confidence interval** for β_{jk} using the following:

$$\hat{\beta}_{jk} \pm z^* SE(\hat{\beta}_{jk})$$

where z^* is calculated from the $N(0,1)$ distribution

> We are $C\%$ confident that for every one unit change in x_j, the predicted odds of $y = k$ versus the baseline $y = 1$ multiply by a factor of $\exp\{\hat{\beta}_{jk} - z^* SE(\hat{\beta}_{jk})\}$ to $\exp\{\hat{\beta}_{jk} + z^* SE(\hat{\beta}_{jk})\}$, holding all else constant.

STA 210

Inference for coefficients

Refer to the model for the NHANES data:

1. Interpret the 95% confidence interval for the coefficient of `Age` in terms of the odds that a person rates themselves as having "Poor" health versus baseline category.

2. Interpret the 95% confidence interval for the coefficient of `PhysActiveYes` in terms of the odds that a person rates themselves as having "Very Good" health versus baseline category.

Predictions

Calculating probabilities

- For categories $k = 2, \dots, K$, the probability that the i^{th} observation is in the j^{th} category is

$$\hat{\pi}_{ij} = \frac{\exp\{\hat{\beta}_{0j} + \hat{\beta}_{1j}x_{i1} + \dots + \hat{\beta}_{pj}x_{ip}\}}{1 + \sum\limits_{k=2}^{K}\exp\{\hat{\beta}_{0k} + \hat{\beta}_{1k}x_{i1} + \dots \hat{\beta}_{pk}x_{ip}\}}$$

- For the baseline category, $k = 1$, we calculate the probability $\hat{\pi}_{i1}$ as

$$\hat{\pi}_{i1} = 1 - \sum_{k=2}^{K}\hat{\pi}_{ik}$$

STA 210

NHANES: Predicted probabilities

```
#calculate predicted probabilities
pred_probs <- as_tibble(predict(health_m, type = "probs")) %>%
                        mutate(obs_num = 1:n())
```

```
pred_probs %>%
  slice(101:105)
```

```
## # A tibble: 5 x 6
##   Excellent Vgood  Good   Fair    Poor obs_num
##       <dbl> <dbl> <dbl>  <dbl>   <dbl>   <int>
## 1    0.0705 0.244 0.451 0.198  0.0366      101
## 2    0.0702 0.244 0.441 0.202  0.0426      102
## 3    0.0696 0.244 0.427 0.206  0.0527      103
## 4    0.0696 0.244 0.427 0.206  0.0527      104
## 5    0.155  0.393 0.359 0.0861 0.00662     105
```

STA 210

Add predictions to original data

```
health_m_aug <- inner_join(nhanes_adult, pred_probs,
                           by = "obs_num") %>%
  select(obs_num, everything())
```

```
health_m_aug %>%
  glimpse()
```

```
## Observations: 6,710
## Variables: 9
## $ obs_num    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
## $ HealthGen  <fct> Good, Good, Good, Good, Vgood, Vgood, Vgood, Vgood, V
## $ Age        <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 50, 33, 60, 5
## $ PhysActive <fct> No, No, No, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No
## $ Excellent  <dbl> 0.07069715, 0.07069715, 0.07069715, 0.07003173, 0.155
## $ Vgood      <dbl> 0.2433979, 0.2433979, 0.2433979, 0.2444214, 0.3922335
## $ Good       <dbl> 0.4573727, 0.4573727, 0.4573727, 0.4372533, 0.3599639
## $ Fair       <dbl> 0.19568909, 0.19568909, 0.19568909, 0.20291032, 0.085
## $ Poor       <dbl> 0.032843150, 0.032843150, 0.032843150, 0.045383332, 0
```

Actual vs. Predicted Health Rating

- We can use our model to predict a person's perceived health rating given their age and whether they exercise

- For each observation, the predicted perceived health rating is the category with the highest predicted probability

```
health_m_aug <- health_m_aug %>%
  mutate(pred_health = predict(health_m, type = "class"))
```

STA 210

Actual vs. Predicted Health Rating

```
health_m_aug %>%
  count(HealthGen, pred_health, .drop = FALSE) %>%
  pivot_wider(names_from = pred_health, values_from = n)
```

```
## # A tibble: 5 x 6
##   HealthGen Excellent Vgood  Good  Fair  Poor
##   <fct>         <int> <int> <int> <int> <int>
## 1 Excellent         0   550   223     0     0
## 2 Vgood             0  1376   785     0     0
## 3 Good              0  1255  1399     0     0
## 4 Fair              0   300   642     0     0
## 5 Poor              0    24   156     0     0
```

```
#rows = actual, columns = predicted
```

Why do you think no observations were predicted to have a rating of "Excellent", "Fair", or "Poor"?

STA 210