Multinomial Logistic Regression

Introduction

Prof. Maria Tackett

STA 210

[Click for PDF of slides]

Topics

- Introduce multinomial logistic regression
- Interpret model coefficients
- Inference for a coefficient β_{jk}

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

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

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 odds of $y = 1$ versus the baseline $y = 0$ are expected to 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 $\exp\{\hat{\beta}_0\}$

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 $\sum_{k=1}^{K} \pi_k = 1$

STA 210

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

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

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

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.

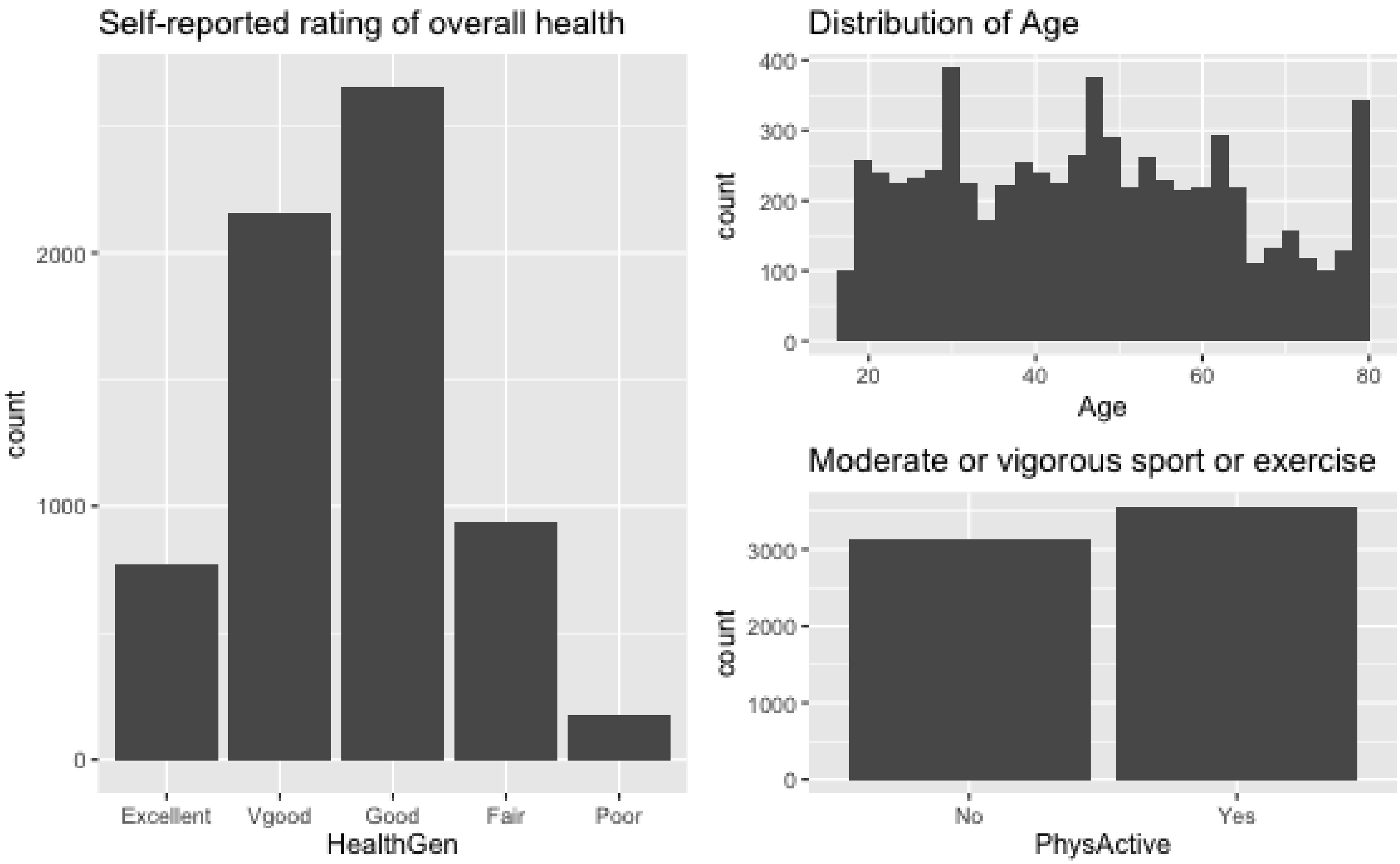
  - **`PhysActive`:** Participant does moderate to vigorous-intensity sports, fitness or recreational activities

The data

```
## Rows: 6,710
## Columns: 4
## $ HealthGen  <fct> Good, Good, Good, Good, Vgood, Vgood, Vgood, Vgood, Vgood, …
## $ Age        <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 50, 33, 60, 56, 56,…
## $ PhysActive <fct> No, No, No, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No, No, …
## $ obs_num    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, …
```

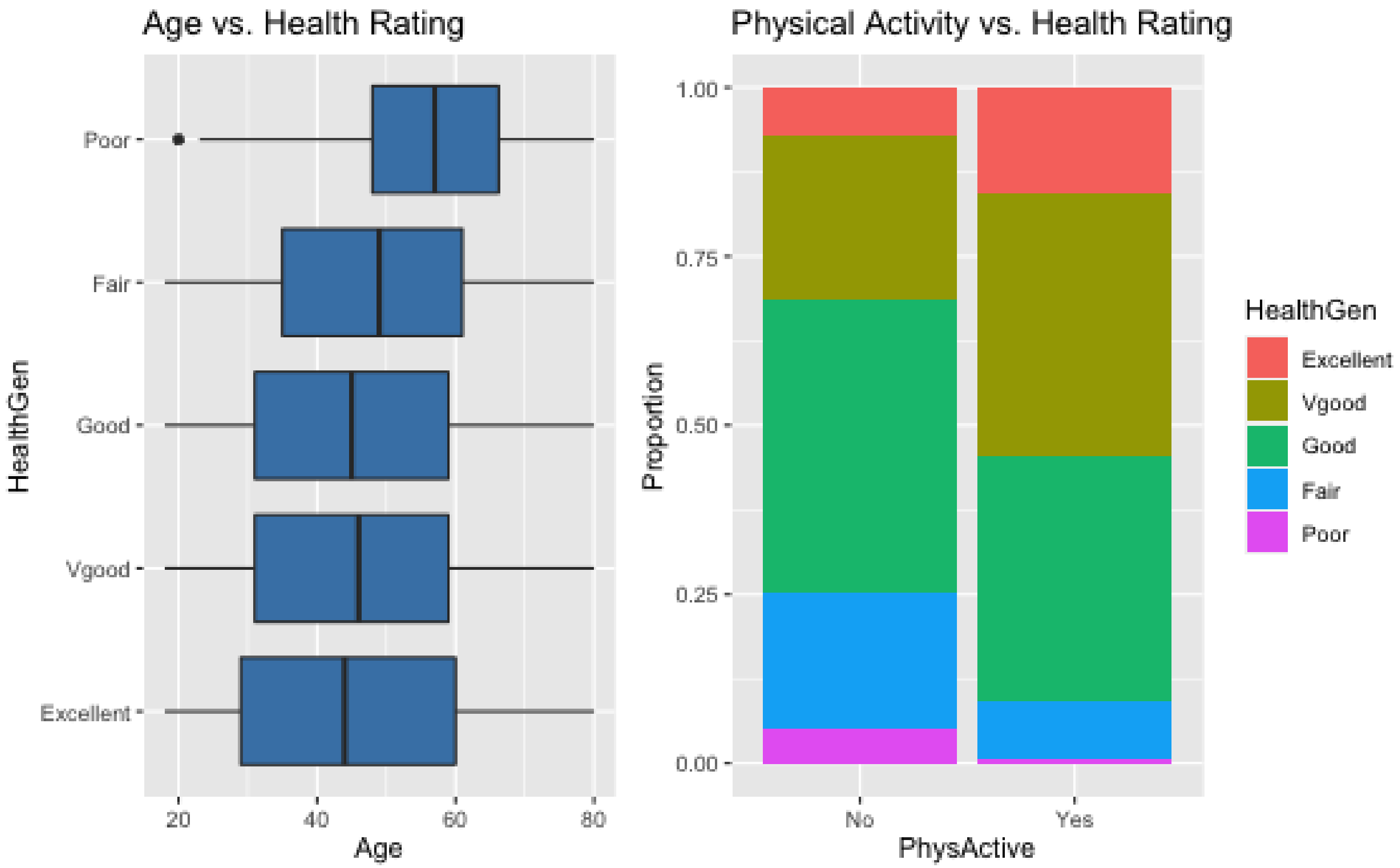
STA 210

Exploratory data analysis



STA 210

Exploratory data analysis



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

Output results

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

STA 210

Model output

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

STA 210

Fair vs. Excellent Health

The baseline category for the model is **`Excellent`**.

The model equation for the log-odds a person rates themselves as having "Fair" health vs. "Excellent" is

$$\log\left(\frac{\hat{\pi}_{Fair}}{\hat{\pi}_{Excellent}}\right) = 0.915 + 0.003 \text{ age} - 1.645 \text{ PhysActive}$$

Interpretations

$$\log\left(\frac{\hat{\pi}_{Fair}}{\hat{\pi}_{Excellent}}\right) = 0.915 + 0.003 \text{ age} - 1.645 \text{ PhysActive}$$

For each additional year in age, the odds a person rates themselves as having fair health versus excellent health are expected to multiply by 1.003 (exp(0.003)), holding physical activity constant.

The odds a person who does physical activity will rate themselves as having fair health versus excellent health are expected to be 0.193 (exp(-1.645)) times the odds for a person who doesn't do physical activity, holding age constant.

Interpretations

$$\log\left(\frac{\hat{\pi}_{Fair}}{\hat{\pi}_{Excellent}}\right) = 0.915 + 0.003 \text{ age} - 1.645 \text{ PhysActive}$$

The odds a 0 year old person who doesn't do physical activity rates themselves as having fair health vs. excellent health are 2.497 (exp(0.915)).

⚠️ **Need to mean-center age for the intercept to have a meaningful interpretation!**

Hypothesis test for β_{jk}

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

STA 210

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 odds of $y = k$ versus the baseline will 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.

Interpreting confidence intervals for β_{jk}

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Fair	(Intercept)	0.915	0.164	5.566	0.00	0.592	1.237
Fair	Age	0.003	0.003	1.058	0.29	-0.003	0.009
Fair	PhysActiveYes	-1.645	0.107	-15.319	0.00	-1.856	-1.435

We are 95% confident, that for each additional year in age, the odds a person rates themselves as having fair health versus excellent health will multiply by 0.997 (exp(-0.003)) to 1.009 (exp(0.009)) , holding physical activity constant.

STA 210

Interpreting confidence intervals for β_{jk}

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Fair	(Intercept)	0.915	0.164	5.566	0.00	0.592	1.237
Fair	Age	0.003	0.003	1.058	0.29	-0.003	0.009
Fair	PhysActiveYes	-1.645	0.107	-15.319	0.00	-1.856	-1.435

We are 95% confident that the odds a person who does physical activity will rate themselves as having fair health versus excellent health are 0.156 (exp(-1.856)) to 0.238 (exp(-1.435)) times the odds for a person who doesn't do physical activity, holding age constant.

Recap

- Introduce multinomial logistic regression
- Interpret model coefficients
- Inference for a coefficient β_{jk}