Introduction to Regression

Logistic Regression

Kara E. McCormack

Welcome!

Topics

- Logistic regression within GLM framework
- Inference: hypothesis tests, confidence intervals, interpretations
- Example with code
- Activity

Slides available here.

Computational Setup

```r
# load packages
library(tidyverse)
library(tidymodels)
library(openintro)
library(knitr)
library(RColorBrewer)
```

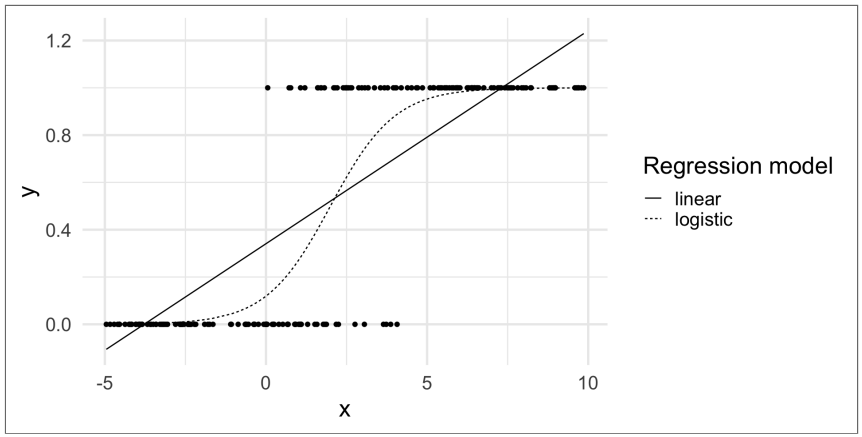
Assumptions

- Familiar with linear regression
- Some familiarity with definition of a generalized linear model (GLM), exponential form, and link function
- Familiar with R and tidyverse
- Interested in reviewing logistic regression + inference

Logistic regression

Linear vs. logistic regression

- Suppose response Y takes value 1 with probability π and value 0 with probability $1 - \pi$.
- Linear regression doesn't fit data well, and produces predicted probabilities below 0 and above 1.
- However, logistic regression always produces probabilities between 0 and 1.



Plot from **BMLR Chapter 6**.

Logistic regression setup

- Suppose response Y takes value 1 with probability π and value 0 with probability $1 - \pi$.
- $\frac{\pi}{1-\pi}$: **odds** that $Y = 1$
- $\log\left(\frac{\pi}{1-\pi}\right)$: **log odds**
- How do we get from π to $\log\left(\frac{\pi}{1-\pi}\right)$? With the **logit transformation**.

Odds to probabilities

- We've seen how to get from probability to odds, now how about odds to probability?

Odds

$$\omega = \frac{\pi}{1 - \pi}$$

Probability

$$\pi = \frac{\omega}{1 + \omega}$$

From odds to probabilities

- **logistic model**: log odds = $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$
- **odds** $= \exp\{\log(\frac{\pi}{1-\pi})\} = \frac{\pi}{1-\pi}$
- Combining this w/ previous slide, we get:

$$\text{probability} = \pi = \frac{\exp\{\beta_0 + \beta_1 X\}}{1 + \exp\{\beta_0 + \beta_1 X\}}$$

Logistic regression: a GLM

- Logistic regression is a **generalized linear model** in which we can analyze data with a dichotomous response with $P(\text{success} = \pi)$.
- **Bernoulli**: Responses are either success $(Y = 1)$ or failure $(Y = 0)$

  $$P(Y = y) = \pi^y (1 - \pi)^{1-y}, \quad y = 0, 1$$

- **Binomial**: Each observation has n bernoulli trials, each with $P(\text{success}) = \pi$.

  $$P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}$$

Binomial or Bernoulli?

1. Is exposure to a particular chemical associated with a cancer diagnosis?
2. Absenteeism data are collected for 146 randomly selected students in New South Wales, Australia across one school year. Are demographic characteristics of children associated with absenteeism?

To submit answers: or click **here**.

01:30

Binomial or Bernoulli?

1. Is exposure to a particular chemical associated with a cancer diagnosis?

- Binomial: The outcome is whether or not a person was diagnosed with cancer.

2. Absenteeism data are collected for 146 randomly selected students in New South Wales, Australia across one school year. Are demographic characteristics of children associated with absenteeism?

   Bernoulli: The outcome is the number of days a student was absent out of n days in a school year.

Exponential form

A bernoulli random variable can be written in one-parameter exponential family form,

$f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right]$

Bernoulli

$$f(y;\pi) = \exp\left[y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right]$$

> What are $a(y), b(\pi), c(\pi)$, and $d(y)$?

01:00

Exponential form

A bernoulli random variable can be written in one-parameter exponential family form,
$f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right]$

Bernoulli

$$f(y;\pi) = \exp\left[y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right]$$

> What are $a(y), b(\pi), c(\pi)$, and $d(y)$?

$a(y) = y$, $b(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, $c(\pi) = \log(1-\pi)$, and $d(y) = 0$.

$b(\pi)$ is the **canonical link function**.

Assumptions of logistic regression

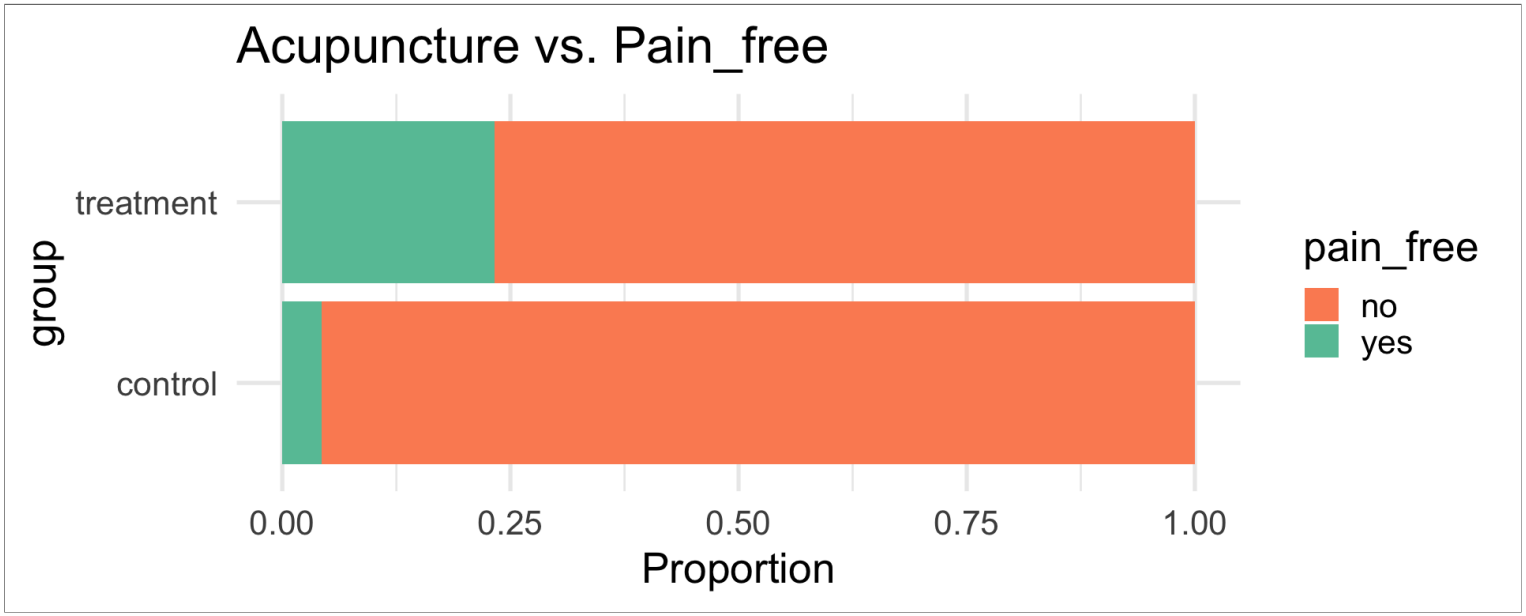
1. **Binary responses**: Response is dichotomous (only takes on two values), or is the sum of dichotomous responses.
2. **Independence**: Observations independent of one another.
3. **Variance structure**: Variance of binomial random variable is $n\pi(1-\pi)$, variance highest when $\pi = 0.5$.
4. **Linearity**: Log of the odds ratio, $\log(\frac{\pi}{1-\pi})$, is a linear function of x.

Acupuncture example

- The `openintro::migraine` dataset is from a study about ear acupuncture in treatment of migraine attacks.
- **Response**: `pain_free` = yes or no
- **Predictor**: `group` = control or treatment
- **Research question**: Is acupuncture treatment associated with a reduction of pain?

Exploratory Data Analysis

- **Research question:** Is acupuncture treatment associated with a reduction of pain?



Modeling being pain-free

```r
acu_model <- glm(pain_free ~ group,
                 data  = migraine,
                 family = "binomial")
acu_model %>%
  tidy %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3.091	0.723	-4.276	0.000
grouptreatment	1.897	0.808	2.348	0.019

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.091 + 1.897 \times \text{treatment}$$

Interpreting treatment coefficient - log odds

term	estimate	std.error	statistic	p.value
(Intercept)	-3.091	0.723	-4.276	0.000
grouptreatment	1.897	0.808	2.348	0.019

The **log-odds** of being pain-free post-treatment are expected to be 1.897 higher for those who received treatment compared to those who did not receive treatment.

Interpreting treatment coefficient - odds

term	estimate	std.error	statistic	p.value
(Intercept)	-3.091	0.723	-4.276	0.000
grouptreatment	1.897	0.808	2.348	0.019

The **odds** of being pain-free post-treatment for those who received treatment are expected to be 6.67 (i.e. exp(1.897)) times the odds for those who received the control.

Hypothesis test for β_j

Hypotheses: $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$

- H_0: There is no linear relationship between the variable of interest and the log-odds of the response.
- H_A: There **is** a linear relationship between the variable of interest and the log-odds of the response.

Hypothesis test for β_j

Hypotheses: $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$

Test statistic:

$$z = \frac{\hat{\beta}_j - 0}{SE_{\hat{\beta}_j}}$$

P-value: $P(|Z| > |z|)$, where $Z \sim N(0, 1)$.

Confidence interval for β_j

Can calculate a **C% confidence interval** for β_j:

$\hat{\beta}_j \pm z^* SE_{\hat{\beta}_j}$

where z^* comes from $N(0, 1)$.

This is an interval for the change in log-odds of the response for a one-unit increase in x_j.

Interpretation in terms of odds

The change in **odds** for every one-unit change in x_j.

$\exp \hat{\beta}_j \pm z^* SE_{\hat{\beta}_j}$

Interpretation: We are C% confident that for every one-unit increase in x_j, the odds multiply by a factor of $\left\{ \exp \hat{\beta}_j - z^* SE_{\hat{\beta}_j} \right\}$ to $\left\{ \exp \hat{\beta}_j + z^* SE_{\hat{\beta}_j} \right\}$, holding all other variables constant.

Let's look at the coefficient for treatment

term	estimate	std.error	statistic	p.value
(Intercept)	-3.091	0.723	-4.276	0.000
grouptreatment	1.897	0.808	2.348	0.019

Test statistic

$$z = \frac{1.897 - 0}{0.808} = 2.34778$$

Let's look at the coefficient for treatment

term	estimate	std.error	statistic	p.value
(Intercept)	-3.091	0.723	-4.276	0.000
grouptreatment	1.897	0.808	2.348	0.019

P-value

$P(|Z| > |2.34778|)$

```
2 * pnorm(2.34778, lower.tail = FALSE)
```

```
[1] 0.01888567
```

Let’s look at the coefficient for treatment

term	est	error	statistic	p.value
(Intercept)		0.723	-4.276	0.000
grouptreatment		0.808	2.348	0.019

Conclusion: Since the p- all, we reject H_0. The data provide sufficient evidence that the acupuncture treatme ly significant predictor of being migraine-pain-free post-treatment.

Multinomial Logistic Regression

Multinomial response

- Suppose our response variable y takes on multiple categories $1, \ldots, K$
- **Multinomial distribution**:

  $P(y = 1) = \pi_1, P(y = 2) = \pi_2, \ldots, P(y = K) = \pi_K$
  with $\sum_{k=1}^{K} \pi_k = 1$

Multinomial logistic regression

- Choose a baseline category for the response (i.e. $y = 1$).

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

- There is a separate equation for each level of response, relative to baseline category.
- If we have K categories of the response, will have $K - 1$ equations as part of ourmultinomial logistic regression model.

NHANES data

- American National Health and Nutrition Examination Survey, NHANES R package, collected by the National Center for Health Statistics (NCHS)
- Survey: Individuals of all ages complete a health exam.
- Data from 2009-2010 and 2011-2012 sample years
- R package data adapted for educational purposes, not suitable for research
- For research purposes, download original files from **NCHS website**
- `?NHANES` in R for list of variables

Self-reported health vs. Age & Sleep Trouble

> **Research question**: Is there an association between age, trouble sleeping, and self-reported health status?

- **HealthGen**: self-reported health rating: Poor, Fair, Good, VGood, or Excellent.
- **Age**: age (years) at time of screening. Participants > 80 recorded as 80.
- **SleepTrouble**: has told doctor that they had trouble sleeping: Yes or No.

The data

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

HealthGen	Age	SleepTrouble	obs_num
Good	34	Yes	1
Good	34	Yes	2
Good	34	Yes	3
Good	49	Yes	4
Vgood	45	No	5
Vgood	45	No	6

Exploratory Data Analysis

Age | Trouble Sleeping | Self-Reported Health

Exploratory data analysis

Age vs. Health rating | Sleep trouble vs. Health rating

Model in R

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

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

- If you don't specify a baseline value of response, R defaults to first level alphabetically (i.e. excellent).

Output results

```
tidy(health_m, conf.int = TRUE, exponentiate = FALSE) %>%
  head(8) %>%
  kable(digits = 3, format = "markdown")
```

y.level	term	estimate	std.error	statistic	p.value	con
Vgood	(Intercept)	0.900	0.117	7.674	0.000	
Vgood	Age	0.002	0.002	0.740	0.460	
Vgood	SleepTroubleYes	0.252	0.111	2.267	0.023	
Good	(Intercept)	1.044	0.115	9.090	0.000	
Good	Age	0.001	0.002	0.311	0.756	
Good	SleepTroubleYes	0.701	0.106	6.592	0.000	
Fair	(Intercept)	-0.442	0.140	-3.145	0.002	
Fair	Age	0.009	0.003	3.108	0.002	

Poor vs. Excellent health

y.level	term	estimate	std.error	statistic	p.value	conf.low	co
Poor	(Intercept)	-3.567	0.288	-12.383	0	-4.132	
Poor	Age	0.031	0.005	6.156	0	0.021	
Poor	SleepTroubleYes	1.669	0.179	9.318	0	1.318	

- Baseline category of health rating is **Excellent**.
- Model equation: the log odds that a person rates themselves “Poor” vs “Excellent” health is

$$\log\left(\frac{\hat{\pi}_{Poor}}{\hat{\pi}_{Excellent}}\right) = -3.567 + 0.031 \cdot \text{Age} + 1.669 \cdot \text{SleepTrouble}$$

Interpretations

$$\log\left(\frac{\hat{\pi}_{Poor}}{\hat{\pi}_{Excellent}}\right) = -3.567 + 0.031 \cdot \text{Age} + 1.669 \cdot \text{SleepTrouble}$$

For each additional year of age, the odds a person rates themselves as having poor health vs. excellent health are expected to multiply by 1.031 (exp(0.031)), holding sleep trouble constant. For those who have trouble sleeping, the odds they rate themselves as having poor health versus excellent health are expected to multiply by 5.306 (exp(1.669)), holding age constant.

Interpretations: intercept

$$\log\left(\frac{\hat{\pi}_{Poor}}{\hat{\pi}_{Excellent}}\right) = -3.567 + 0.031 \cdot \text{Age} + 1.669 \cdot \text{SleepTrouble}$$

What is the interpretation for the intercept of this model, in terms of odds?

Interpretations: intercept

$$\log\left(\frac{\hat{\pi}_{Poor}}{\hat{\pi}_{Excellent}}\right) = -3.567 + 0.031 \cdot \text{Age} + 1.669 \cdot \text{SleepTrouble}$$

The odds a 0 year-old person without sleep trouble rates themselves as having poor health versus excellent health are 0.028 (exp(-3.567)).

- Would need to mean-center age for the intercept to have a meaningful interpretation.

Confidence Interval for Sleep Trouble

y.level	term	estimate	p.value	conf.low	conf.high
Poor	(Intercept)	-3.567	0	-4.132	-3.003
Poor	Age	0.031	0	0.021	0.040
Poor	SleepTroubleYes	1.669	0	1.318	2.020

We are 95% confident that, if someone has trouble sleeping, the odds the person rates themselves as poor health vs excellent health will multiply by 3.735 (exp(1.318)) to 7.538 (exp(2.020)), holding age constant.

Visualization: forest plots

```r
model_coef <- tidy(health_m, exponentiate = TRUE, conf.int = TRUE) %>%
  filter(y.level %in% c("Fair", "Poor"))
ggplot(data = model_coef, aes(x = term, y = estimate)) +
  geom_point() +
  geom_hline(yintercept = 1, lty = 2) +
  geom_pointrange(aes(ymin = conf.low, ymax = conf.high))+
  labs(title = "Exponentiated model coefficients") +
  coord_flip() +
  facet_wrap(~y.level)
```

Recap

- Logistic regression in context of GLM
- Multinomial logistic regression + inference, health rating example
- Next up, activity

Activity

Regression Bingo Game

- Pair up - two people per bingo card.
- Each square on bingo card has a question.
- “Answers” located throughout room. If you think you’ve found a correct answer, take a sticker and place it in the square.
  - Write a note on your card about what the answer said
- When you get bingo (3 in a row), shout it out and share your 3 question/answers.
- If you’d like to see any slide from this lecture, feel free to ask!

Acknowledgements

- BMLR Chapter 6
- Introduction to Modern Statistics, Chapter 9
- STA210: Regression Analysis

That’s all, folks!