Logistic regression

02.17.22

Click here for PDF of slides

Announcements

- Fill out mini-project 01 team evaluation by **TODAY at 11:59pm**
- Quiz 02 on Gradescope **due Fri, Feb 18 at 11:59pm**

Learning goals

- Identify Bernoulli and binomial random variables
- Write GLM for binomial response variable
- Interpret the coefficients for a logistic regression model
- Visualizations for logistic regression

Basics of logistic regression

Bernoulli + Binomial random variables

Logistic regression is used to analyze data with two types of responses:

- **Binary**: These responses take on two values success $(Y = 1)$ or failure $(Y = 0)$, yes $(Y = 1)$ or no $(Y = 0)$, etc.

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

- **Binomial**: Number of successes in a Bernoulli process, n independent trials with a constant probability of success p.

$$P(Y = y) = \binom{n}{y} p^y(1 - p)^{n-y} \quad y = 0, 1, \ldots, n$$

In both instances, the goal is to model p the probability of success.

Binary vs. Binomial data

For each example, identify if the response is a Bernoulli or Binomial response:

1. Use median age and unemployment rate in a county to predict the percent of Obama votes in the county in the 2008 presidential election.

2. Use GPA and MCAT scores to estimate the probability a student is accepted into medical school.

3. Use sex, age, and smoking history to estimate the probability an individual has lung cancer.

4. Use offensive and defensive statistics from the 2017-2018 NBA season to predict a team's winning percentage.

Click here to submit your responses.

02:30

Logistic regression model

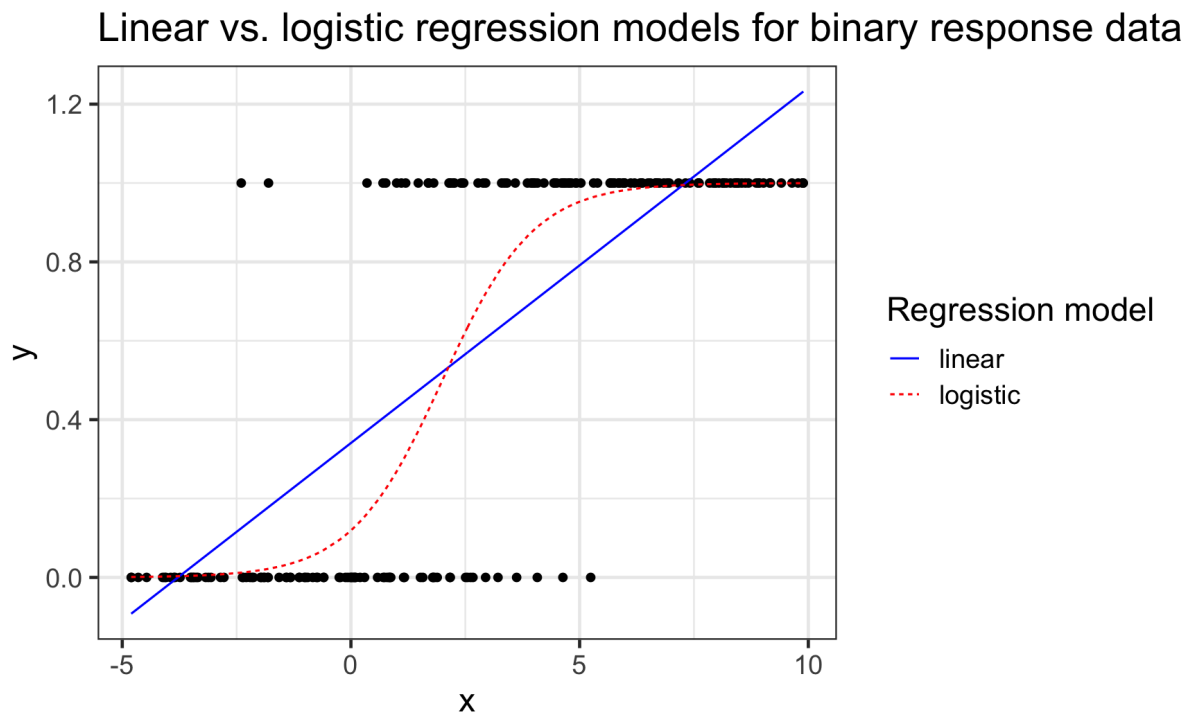
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- The response variable, $\log\left(\frac{p}{1-p}\right)$, is the log(odds) of success, i.e. the logit
- Use the model to calculate the probability of success

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

- When the response is a Bernoulli random variable, the probabilities can be used to classify each observation as a success or failure

Logistic vs linear regression model



Graph from BMLR Chapter 6

Logit link

Bernoulli and Binomial random variables can be written in one-parameter exponential family form, $f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$

Bernoulli

$$f(y;p) = e^{y\log(\frac{p}{1-p})+\log(1-p)}$$

Binomial

$$f(y;n,p) = e^{y\log(\frac{p}{1-p})+n\log(1-p)+\log\binom{n}{y}}$$

They have the same canonical link $b(p) = \log\left(\frac{p}{1-p}\right)$

Assumptions for logistic regression

The following assumptions need to be satisfied to use logistic regression to make inferences

1. **Binary response**: The response is dichotomous (has two possible outcomes) or is the sum of dichotomous responses

2. **Independence**: The observations must be independent of one another

3. **Variance structure**: Variance of a binomial random variable is $np(1-p)$ $(n = 1 \text{ for Bernoulli})$, so the variability is highest when $p = 0.5$

4. **Linearity**: The log of the odds ratio, $\log\left(\frac{p}{1-p}\right)$, must be a linear function of the predictors $x_1, \ldots, x_p$

COVID-19 infection prevention practices at food establishments

Researchers at Wollo Univeristy in Ethiopia conducted a study in July and August 2020 to understand factors associated with good COVID-19 infection prevention practices at food establishments. Their study is published in Andualem et al. (2022)

They were particularly interested in the understanding implementation of prevention practices at food establishments, given the workers' increased risk due to daily contact with customers.

Andualem, A., Tegegne, B., Ademe, S., Natnael, T., Berihun, G., Abebe, M., ... & Adane, M. (2022). COVID-19 infection prevention practices among a sample of food handlers of food and drink establishments in Ethiopia. PloS one, 17(1), e0259851.

The data

*"An institution-based cross-sectional study was conducted among **422 food handlers in Dessie City and Kombolcha Town food and drink establishments in July and August 2020.** The study participants were selected using a **simple random sampling** technique. Data were collected by trained data collectors using a pretested structured questionnaire and an on-the-spot observational checklist."*

Andualem, A., Tegegne, B., Ademe, S., Natnael, T., Berihun, G., Abebe, M., ... & Adane, M. (2022). COVID-19 infection prevention practices among a sample of food handlers of food and drink establishments in Ethiopia. PloS one, 17(1), e0259851.

Response variable

*"The outcome variable of this study was the **good or poor practices of COVID-19 infection prevention among food handlers**. Nine yes/no questions, one observational checklist and five multiple choice infection prevention practices questions were asked with a minimum score of 1 and maximum score of 25. Good infection prevention practice (the variable of interest) was determined for food handlers who scored 75% or above, whereas poor infection prevention practices refers to those food handlers who scored below 75% on the practice questions."*

Andualem, A., Tegegne, B., Ademe, S., Natnael, T., Berihun, G., Abebe, M., ... & Adane, M. (2022). COVID-19 infection prevention practices among a sample of food handlers of food and drink establishments in Ethiopia. PloS one, 17(1), e0259851.

Results

Variables		Infection prevention practices status		AOR (95% CI)	*P*-value
		Good	Poor		
		n	n		
Educational status	Unable to read and write	7	12	Ref	
	Informal education	10	9	1.32(0.50–2.10)	0.338
	Primary school	36	29	1.56(0.86–2.37)	0.956
	Secondary school	141	66	2.40(1.22–4.73)	0.543
	College or above	60	31	1.97(1.32–3.75)	0.042
Years of experience	<1	53	104	0.84(0.51–1.38)	0.678
	1–5	85	111	Ref	
	>5	38	10	2.55(1.43–5.77)	0.025
Availability of COVID-19 infection prevention guidelines in food and drink establishment	Yes	66	30	2.68(1.52–4.75)	< 0.001
	No	110	195	Ref	
Ever taken COVID-19 infection prevention training	Yes	44	14	3.26(1.61–6.61)	< 0.001
	No	132	211	Ref	

Ref, reference category; AOR, adjusted odds ratio

https://doi.org/10.1371/journal.pone.0259851.t007

Andualem, A., Tegegne, B., Ademe, S., Natnael, T., Berihun, G., Abebe, M., ... & Adane, M. (2022). COVID-19 infection prevention practices among a sample of food handlers of food and drink establishments in Ethiopia. PloS one, 17(1), e0259851.

Interpreting the results

- Is the response a Bernoulli or Binomial?
- What is the strongest predictor of having good COVID-19 infection prevention practices?
    - It's often unreliable to look answer this question just based on the model output. Why are we able to answer this question based on the model output in this case?
- Describe the effect (coefficient interpretation and inference) of having COVID-19 infection prevention policies available at the food establishment.
- The intercept describes what group of food handlers?

04:30

Visualizations for logistic regression

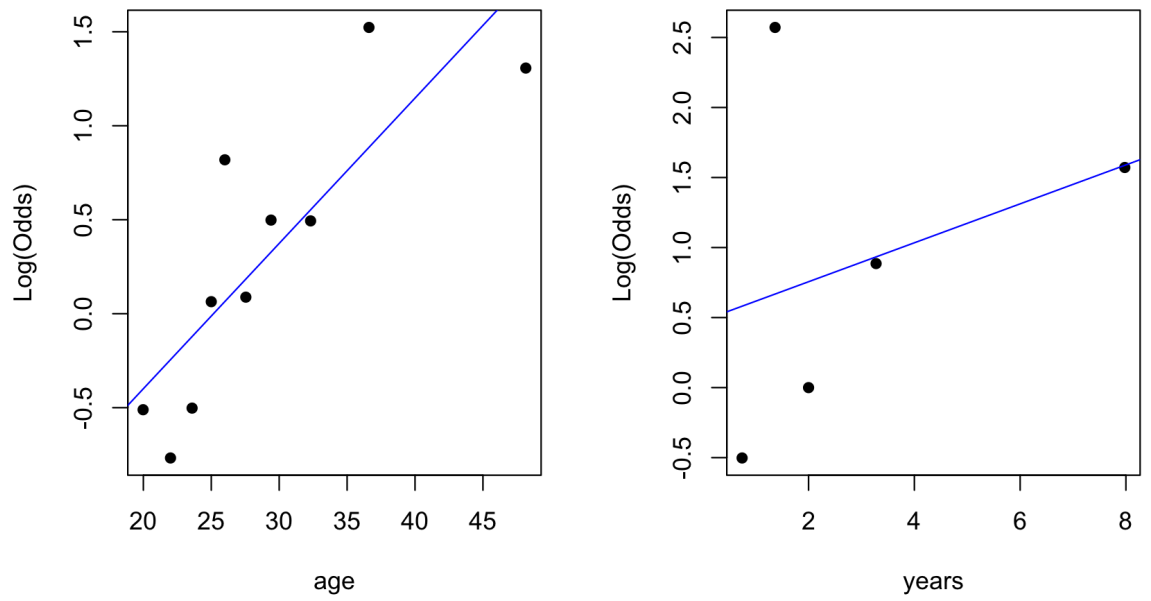
Access to personal protective equipment

We will use the data from Andualem et al. (2022) to explore the association between age, sex, years of service, and whether someone works at a food establishment with access to personal protective equipment (PPE) as of August 2020. We will use access to PPE as a proxy for wearing PPE.

age	sex	years	ppe_access
34	Male	2	1
32	Female	3	1
32	Female	1	1
40	Male	4	1
32	Male	10	1

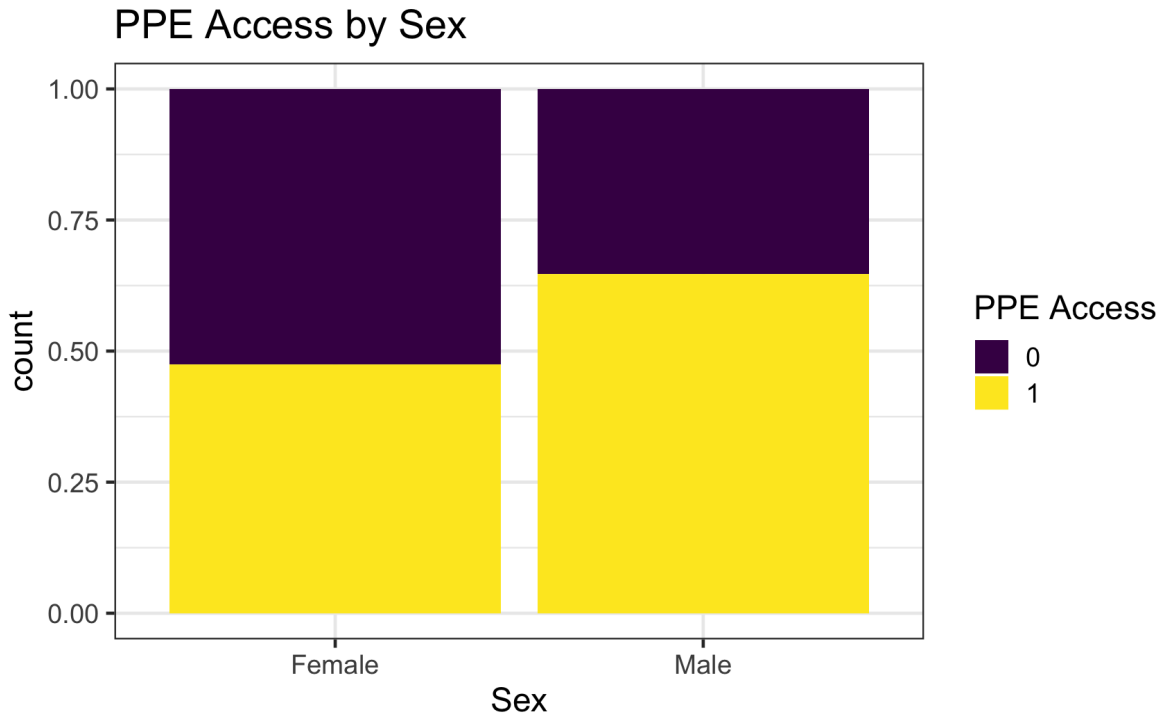
Exploratory data analysis for binary response

```
library(Stat2Data)
par(mfrow = c(1, 2))
emplogitplot1(ppe_access ~ age, data = covid_df, ngroups = 10)
emplogitplot1(ppe_access ~ years, data = covid_df, ngroups = 5)
```



Exploratory data analysis for binary response

```
library(viridis)
ggplot(data = covid_df, aes(x = sex, fill = factor(ppe_access))) +
  geom_bar(position = "fill")  +
  labs(x = "Sex",
       fill = "PPE Access",
       title = "PPE Access by Sex") +
  scale_fill_viridis_d()
```



Model results

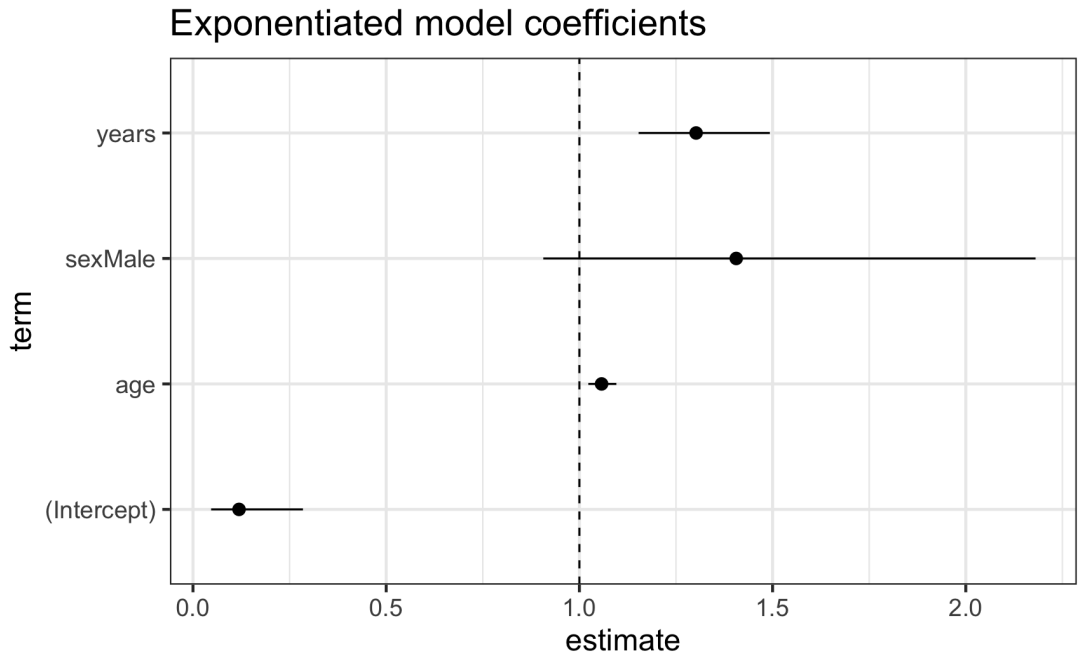
```
ppe_model <- glm(factor(ppe_access) ~ age + sex + years, data = covid_df,
                 family = binomial)
tidy(ppe_model, conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.127	0.458	-4.641	0.000	-3.058	-1.257
age	0.056	0.017	3.210	0.001	0.023	0.091
sexMale	0.341	0.224	1.524	0.128	-0.098	0.780
years	0.264	0.066	4.010	0.000	0.143	0.401

Visualizing coefficient estimates

```
model_coef <- tidy(ppe_model, exponentiate = TRUE, conf.int = TRUE)
```

```
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()
```



Acknowledgements

These slides are based on content in

- BMLR: Chapter 6 - Logistic Regression