Lesson 16: Log-binomial Regression

Nicky Wakim

2025-05-28

Learning Objectives

1. Revisit the key distinctions between logistic and log-binomial regression.
2. Interpret and visualize coefficient estimates, risk ratios, and predicted probabilities.

Learning Objectives

1. Revisit the key distinctions between logistic and log-binomial regression.
2. Interpret and visualize coefficient estimates, risk ratios, and predicted probabilities.

Log-binomial Regression

- **Outcome type:** binary, yes or no

- **Example outcomes:**
  - Food insecurity
  - Disease diagnosis for patient
  - Fracture

- **Population model**

$\mu = P(Y = 1 \mid X)$

$$\log(\mu) = \beta_0 + \beta_1 X$$

- **Interpretations**
  - We have log of probability on the left
  - So exponential of our coefficients will be **risk ratio**

Logistic regression vs. log-binomial regression

- Outcome is the same: still a binary variable
  - Still modeling with probability of event: $P(Y = 1|X)$
- Logistic = logit = log of odds
  - Aim to get **odds ratios** for interpretation (exponential of coefficients)
  - You *can* convert the odds ratios to risk ratios
- Log-binomial = log of probability
  - Aim to get **risk ratio** or prevalence ratio (exponential of coefficients)

rare events
OR ≈ RR

OR → RR = $\frac{\hat{\pi}_1}{\hat{\pi}_2}$

predicted prob in logistic

logit($\pi(X)$)

log($\pi(X)$)

harder b/c bounded

log($\pi(X)$) ∈ $[0, \infty)$

- Disadvantage of log-binomial
  - Left hand side (aka $\log(\pi(X))$) is only a positive value, while right hand side can range from $-\infty$ to ∞
  - Prone to issues while trying to fit the model!

logit($\pi(X)$) ∈ $(-\infty, \infty)$

easier

A few classes ago: GLOW Study with interactions

- **Outcome variable:** any fracture in the first year of follow up (FRACTURE: 0 or 1)
- **Risk factor/variable of interest:** history of prior fracture (PRIORFRAC: 0 or 1)
- **Potential confounder or effect modifier:** age (AGE, a continuous variable)

- Fit a **logistic regression** model with interactions:

$$\text{logit}\left(\pi(\mathbf{X})\right) = \beta_0 + \beta_1 \cdot I(\text{PF} = \text{“Yes”}) + \beta_2 \cdot Age + \beta_3 \cdot I(\text{PF} = \text{“Yes”}) \cdot Age$$

population

$$\text{logit}\left(\widehat{\pi}(\mathbf{X})\right) = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{PF} = \text{“Yes”}) + \widehat{\beta}_2 \cdot Age + \widehat{\beta}_3 \cdot I(\text{PF} = \text{“Yes”}) \cdot Age$$

fitted/est model

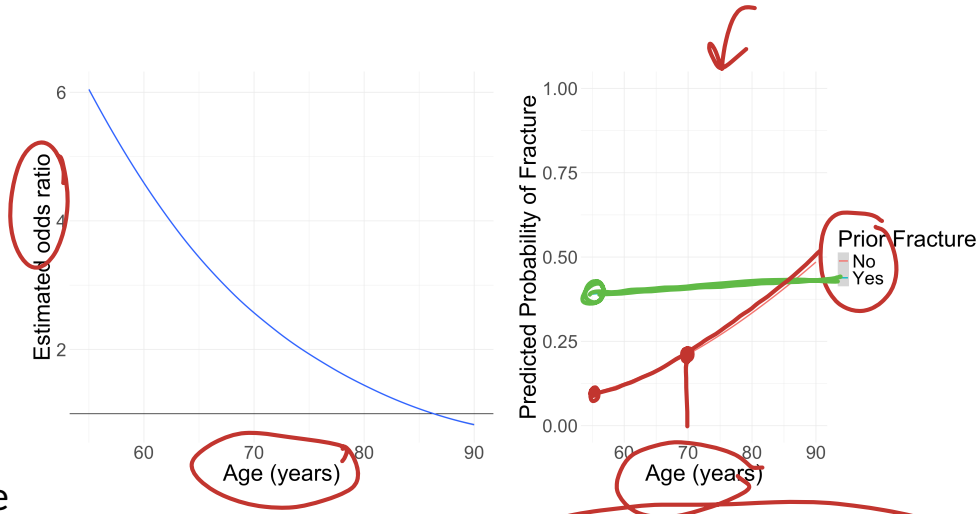
$$\text{logit}\left(\widehat{\pi}(\mathbf{X})\right) = -1.376 + 1.002 \cdot I(\text{PF} = \text{“Yes”}) + 0.063 \cdot Age - 0.057 \cdot I(\text{PF} = \text{“Yes”}) \cdot Age$$

↳ fitted but w/ values inserted for $\hat{\beta}$s

Logistic regression: Reporting results of GLOW Study with interactions

- Remember our main covariate is prior fracture, so we want to focuse on how age changes the relationship between prior fracture and a new fracture!

For individuals 69 years old, the estimated odds of a new fracture for individuals with prior fracture is 2.72 times the estimated odds of a new fracture for individuals with no prior fracture (95% CI: 1.70, 4.35). As seen in Figure 1 (a), the odds ratio of a new fracture when comparing prior fracture status decreases with age, indicating that the effect of prior fractures on new fractures decreases as individuals get older. In Figure 1 (b), it is evident that for both prior fracture statuses, the predict probability of a new fracture increases as age increases. However, the predicted probability of new fracture for those without a prior fracture increases at a higher rate than that of individuals with a prior fracture. Thus, the predicted probabilities of a new fracture converge at age [insert age here].



(a) Odds ratio of fracture outcome comparing prior fracture to no prior fracture

(b) Predicted probability of fracture

Figure 1: Plots of odds ratio and predicted probability from fitted interaction model

What would a log-binomial equivalent look like?

- For the same GLOW data, we could fit a log-binomial model:

$$\log\left(\widehat{\pi}(\mathbf{X})\right) = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{PF} = \text{“Yes”}) + \widehat{\beta}_2 \cdot Age + \widehat{\beta}_3 \cdot I(\text{PF} = \text{“Yes”}) \cdot Age$$

- What does this look like in the `glm` function?

```
1 glow_log_binom = glow %>% glm(formula = fracture ~ priorfrac + age_c + priorfrac*age_c,
2                               family = binomial(link = "log"))
```

▶ Table for coefficient estimates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	−1.625	0.106	−15.269	0.000	−1.846	−1.427
priorfracYes	0.727	0.159	4.570	0.000	0.404	1.034
age_c	0.046	0.011	4.217	0.000	0.025	0.066
priorfracYes:age_c	−0.043	0.016	−2.710	0.007	−0.074	−0.012

Let's start with a model with main effects only

- Just so we have a simpler model as we first demostrate the log-binomial:

$$\log\left(\widehat{\pi}(\mathbf{X})\right) = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{PF} = \text{“Yes”}) + \widehat{\beta}_2 \cdot Age$$

- Run a logistic regression model using `glm()`: *works!*

```
glow2_logistic = glow %>%
  glm(formula = fracture ~ priorfrac + age_c,
                           family = binomial)
```

- Run a log-binomial model using `glm()`: *does not work!*

```
glow2_log_binom = glow %>%
  glm(formula = fracture ~ priorfrac + age_c,
                           family = binomial(link = "log"))
```

```
Error: no valid set of coefficients has been found: please supply starting values
```

How to fix this?

- We can fit the log-binomial model using `logbin()`
  - This function uses something called the EM algorithm to fit the data
  - Recall that `glm` uses an iteratively reweighted least squares algorithm

in logbin pkg in R

computer' method

```
glow2_log_binom = logbin(formula = fracture ~ priorfrac + age_c,
                         data = glow)
```

- Using `logbin` instead of `glm` will often fix the issue with fitting
  - But it does not work 100% of the time!

▶ Table for coefficient estimates from logbin()

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	−1.595	0.103	−15.428	0.000	−1.798	−1.393
priorfracYes	0.545	0.159	3.416	0.001	0.232	0.857
age_c	0.026	0.008	3.104	0.002	0.009	0.042

Model tests: same as logistic regression

- We would go through the same procedure as logistic regression
  - Because both models are rooted in the likelihood function!

- Use LRT to test multiple variables OR one categorical variable with multiple levels

- Use Wald test to construct confidence intervals
  - Or test one continuous or one binary variable

Learning Objectives

1. Revisit the key distinctions between logistic and log-binomial regression.

2. Interpret and visualize coefficient estimates, risk ratios, and predicted probabilities.

Interpreting risk ratios from log-binomial regression

- Let's say we ran the following model:

$$\log\left(\hat{\pi}(\mathbf{X})\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot I(\text{PF} = \text{“Yes”}) + \hat{\beta}_2 \cdot Age$$

centered age

estimated

- $\exp(\hat{\beta}_0)$: The risk of fracture (outcome event) for someone who had no prior fracture and is the mean age of 69 years old
- $\exp(\hat{\beta}_1)$: The risk of fracture for someone who had a prior fracture is $\exp(\hat{\beta}_1)$ times the risk of fracture for someone who did not have a prior fracture, adjusting for age.
- $\exp(\hat{\beta}_2)$: For every one year increase in age, the risk of fracture is $\exp(\hat{\beta}_2)$ times, adjusting for prior fracture.

estimated

estimated

estimated

*Recall that once I estimate the coefficients, my interpretations need to reflect that they are estimated by saying "estimated risk"

$$\text{Risk Ratio} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{\text{Risk}_1}{\text{Risk}_2}$$

Interpretation of risk ratio: no interactions (1/2)

- If we have no interactions in the model (using `logbin`):

```
glow_main = logbin(formula = fracture ~ priorfrac + age_c, data = glow)
```

- Look at the exponential of the coefficients (risk ratio):

```
tidy(glow_main, conf.int = T, exponentiate = T) %>%
  gt() %>%
  tab_options(table.font.size = 35) %>%
  fmt_number(decimals = 3)
```

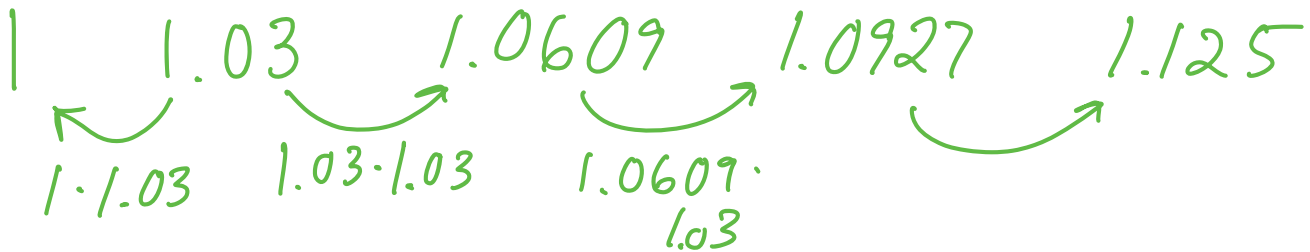
RR estimates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.203	0.103	−15.428	0.000	0.166	0.248
priorfracYes	1.724	0.159	3.416	0.001	1.261	2.356
age_c	1.026	0.008	3.104	0.002	1.010	1.043

Interpretation of risk ratio: no interactions (2/2)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.203	0.103	−15.428	0.000	0.166	0.248
priorfracYes	1.724	0.159	3.416	0.001	1.261	2.356
age_c	1.026	0.008	3.104	0.002	1.010	1.043

- $\exp(\widehat{\beta}_0)$: The estimated risk of fracture is 0.2 for someone who had no prior fracture and is the mean age of 69 years old (95% CI: 0.166, 0.248)
- $\exp(\widehat{\beta}_1)$: The estimated risk of fracture for someone who had a prior fracture is 1.72 times the estimated risk of fracture for someone who did not have a prior fracture, adjusting for age.
- $\exp(\widehat{\beta}_2)$: For every one year increase in age, the estimated risk of fracture is 1.03 times, adjusting for prior fracture.
  - OR: For every one year increase in age, the estimated risk of fracture increases 2.61%, adjusting for prior fracture.



Visualization: We can look at the log-risk

- This will help us understand the model that we fit
- Ultimately, we do not need a visualization of the risk ratio because the main effects model is constant across age

▶ Table of coefficient estimates

term	estimate	std.error	statistic	p.value
(Intercept)	−1.595	0.103	−15.428	0.000
priorfracYes	0.545	0.159	3.416	0.001
age_c	0.026	0.008	3.104	0.002

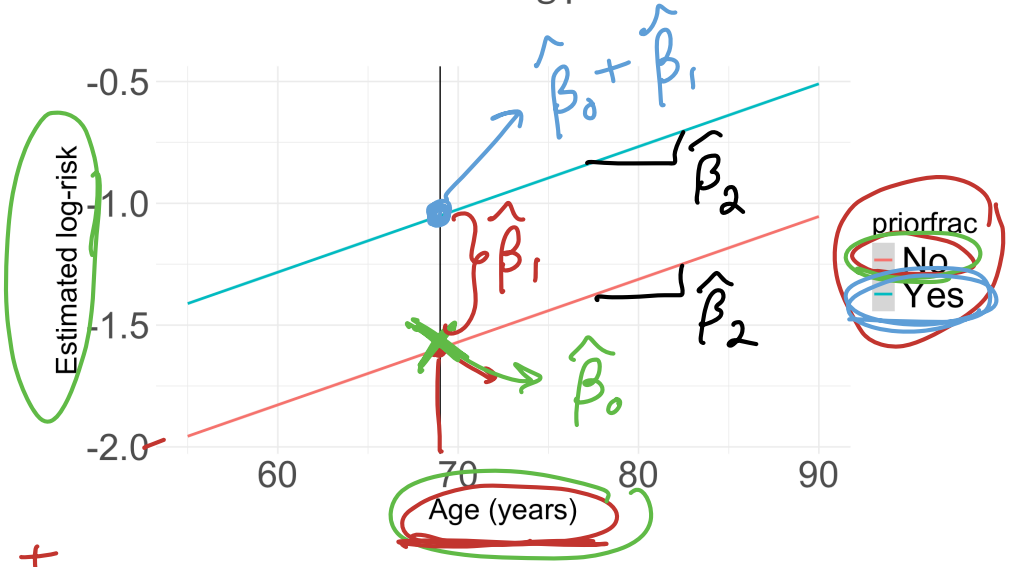
$\log(\hat{\pi}(x)) =$

$-1.595 +$

$0.545\, I(PF = \text{"Yes"}) +$ (= 0)

$0.026\, Age_c$ (= 0)

▶ Code to make the following plot



Interactions: interpretations and visualizations

$$\log\left(\widehat{\pi}(\mathbf{X})\right) = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{PF} = \text{“Yes”}) + \widehat{\beta}_2 \cdot Age + \widehat{\beta}_3 \cdot I(\text{PF} = \text{“Yes”}) \cdot Age$$

```
glow3 = glow %>%
  mutate(interaction_PF_age = as.numeric(priorfrac)*age_c)

glow2_log_binom = logbin(formula = fracture ~ priorfrac + age_c + interaction_PF_age,
          data = glow3)

tidy(glow2_log_binom, conf.int = T) %>%
  gt() %>%
  tab_options(table.font.size = 30) %>%
  fmt_number(decimals = 3)
```

priorfrac x age

coef estimates

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	−1.596	0.104	−15.414	0.000	−1.799	−1.393
priorfracYes	0.576	0.170	3.382	0.001	0.242	0.910
age_c	0.037	0.026	1.440	0.150	−0.013	0.088
interaction_PF_age	−0.009	0.017	−0.546	0.585	−0.042	0.023

Visualization of interaction: We can look at the log-risk

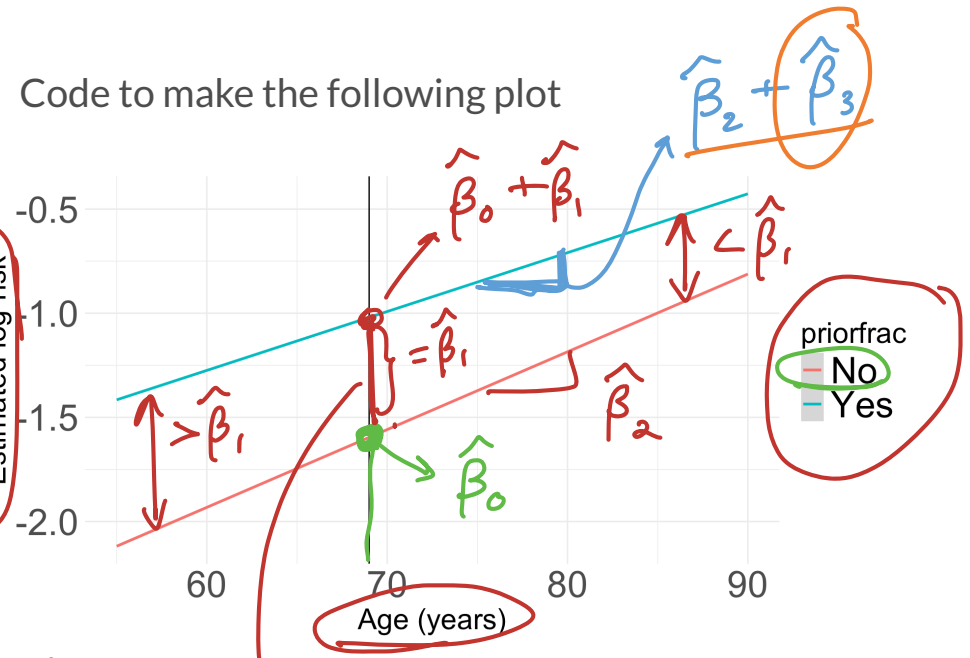
- This will help us understand the model that we fit
- Ultimately, we do not need a visualization of the risk ratio because the main effects model is constant across age

▶ Table of coefficient estimates

term	estimate	std.error	statistic	p.value
(Intercept)	−1.596	0.104	−15.414	0.000
priorfracYes	0.576	0.170	3.382	0.001
age_c	0.037	0.026	1.440	0.150
interaction_PF_age	−0.009	0.017	−0.546	0.585

$\log(\hat{\pi}(x)) = -1.596 + 0.576\, I(PF = \text{"yes"}) + 0.037\, Age_c - 0.009\, I(PF = \text{"yes"}) \cdot Age_c$

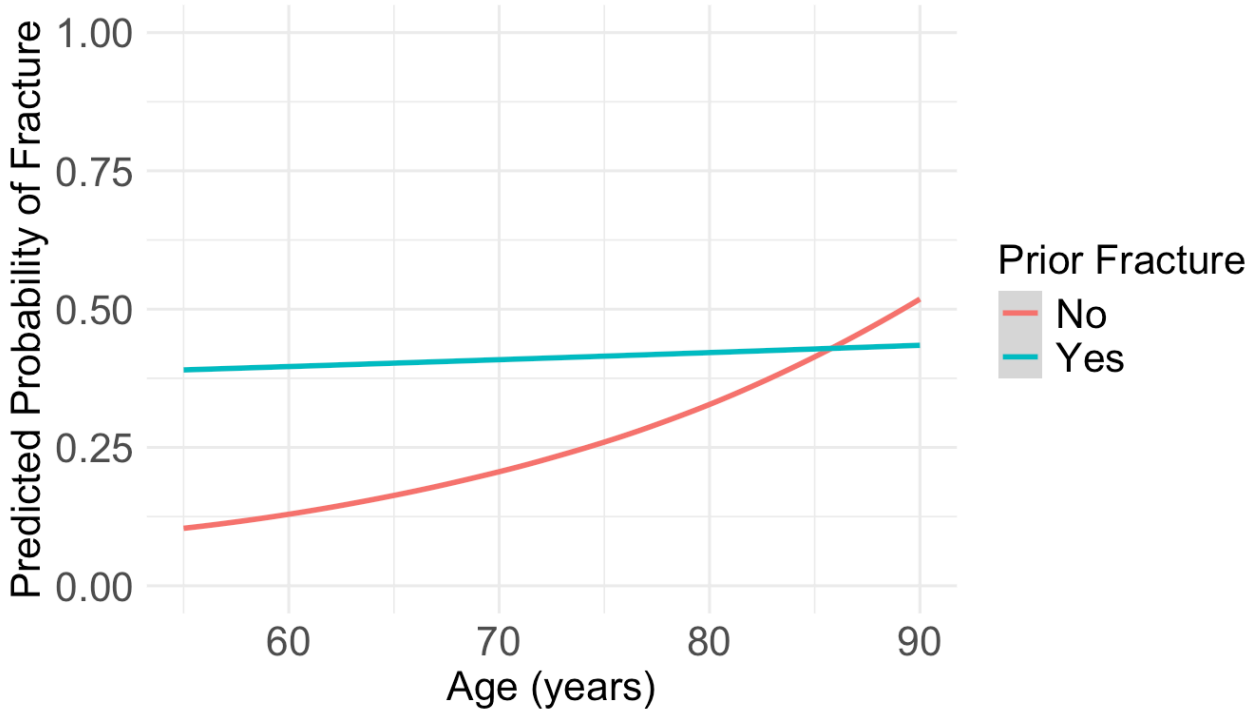
▶ Code to make the following plot



Predicted probabilities and visualizations

- Does this look familiar?? It should!
  - The predicted probabilities from the logistic regression and the log-binomial regression should be the same!

▶ Code to make the following plot



Final words on log-binomial vs logistic

- Log-binomial is more common in field of Epidemiology

- I prefer to use logistic regression because it is more stable AND I can calculate risk ratio from my odds ratio and predicted probability

- If you use $OR \approx RR$, then logistic regression will over-estimate the RR
  - BUT you can still use logistic regression to calculate the RR
  - Find the predicted probabilities for two groups you are comparing and then calculate the RR

$$RR = \frac{\hat{\pi}_1}{\hat{\pi}_2}$$

More resources

- Log-binomial models: exploring failed convergence
- Parameter estimation and goodness-of-fit in log binomial regression
  - Needs PSU login for full access