PHS2000A
Modeling Odds vs. Risks: Logistic and Log Binomial Regression

Jarvis T. Chen (jarvis@hsph.harvard.edu)

20 October 2025

Harvard T. H. Chan School of Public Health

Learning Objectives

- Review logistic regression model and how to calculate odds, odds ratios, and probabilities from output.
- Introduce the log binomial regression model as a technique for modeling risks and risk ratios.
- Discuss the log Poisson model as an alternative to the log binomial model when the latter is difficult to fit.
- Discuss the pros and cons of modeling the log odds vs. the log risk.

Today's Example

To motivate our work today, let's say that we are interested in looking at obesity ($BMI \geq 30$) among White Non-Hispanic and Black Non-Hispanic adults age 25-65 in the NHANES 2007-2008 dataset. Age (in years) and gender are also available in the dataset.

You'll need to load a couple of packages for today.

```
if (!require(sandwich)) install.packages(sandwich)
if (!require(ResourceSelection)) install.packages(ResourceSelection)

# Read in necessary libraries
library(sandwich)
library(ResourceSelection)
```

To read the dataset into R:

```
# Read in dataset
# Make sure the data file is in your working directory.
d.nhanes <- read.csv("lab1_nhanes0708.csv")
```

Some data cleaning steps

```
# Data cleaning steps

# Create BMI variable (weight in kg / height in meters squared)
# Create obesity variable (bmi>=30)
# Create an indicator for black
d.nhanes$bmi <- d.nhanes$weightkg /(d.nhanes$heightcm/100) ^2
d.nhanes$obese <- d.nhanes$bmi>=30
d.nhanes$black <- d.nhanes$race=="bnh"
d.nhanes$female <- d.nhanes$gender==2

# Data cleaning steps
# Drop observations with missing values for bmi and age<25 or age>65
d.obese <- subset(d.nhanes,ageyrs>=25 & ageyrs<65 & !is.na(bmi))

# Center the age variable by subtracting the mean
d.obese$agectr <- d.obese$ageyrs - mean(d.obese$ageyrs)
```

Review of logistic regression

Logistic regression

Let $Y_i \in \{0, 1\}$ represent obesity, x_1 represent age in years (centered), x_2 be a binary indicator for black race/ethnicity, and x_3 be a binary indicator for gender.

Let's begin by fitting a logistic regression model for the log odds of being obese given age, race/ethnicity, and gender.

$$\text{logit}\left[\pi(\mathbf{x}_i)\right] = \log\left[\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

Quick review: We learned that the logistic regression model is a type of **generalized linear model**:

- What is the *systematic* component of this model?
  - $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$
- What is the *random* component of this model?
  - $Y|\mathbf{X} \sim \text{Bernoulli}[\pi(\mathbf{X}_i)]$
- What is the *link function*?
  - $\eta_i = \text{logit}\left[\pi(\mathbf{x}_i)\right]$

Logistic regression

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \text{agectr} + \beta_2 \text{black} + \beta_3 \text{gender}$$

```
logistic.m1 <- glm(obese ~ agectr + black + factor(gender),
                   data=d.obese, family=binomial(link="logit"))
summary(logistic.m1)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.6927     0.0677  -10.23  < 2e-16 ***
agectr            0.0111     0.0037    3.00   0.0027 **
blackTRUE         0.3550     0.0882    4.03  5.7e-05 ***
femaleTRUE        0.2099     0.0843    2.49   0.0128 *
```

Recall in our notation that $\pi(x) = \mathbb{E}(Y = 1|x) = \mu(x)$.

Logistic regression

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \text{agectr} + \beta_2 \text{black} + \beta_3 \text{gender}$$

How do we interpret the $\hat{\beta}$ coefficients from this model?

$\hat{\beta}$	Estimate	Interpretation
$\hat{\beta}_0$	-0.6927	Log odds of obesity at the mean of age, for white non-Hispanic men
$\hat{\beta}_1$	0.0111	Log odds ratio of obesity for a one-unit increase in age, holding race/ethnicity and gender constant
$\hat{\beta}_2$	0.3550	Log odds ratio of obesity for Black Non-Hispanic vs. white non-Hispanic, holding age and gender constant
$\hat{\beta}_3$	0.2099	Log odds ratio of obesity for women vs. men, holding age and race/ethnicity constant

Logistic regression

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \text{agectr} + \beta_2 \text{black} + \beta_3 \text{gender}$$

How do we interpret $e^{\hat{\beta}}$?

$e^{\hat{\beta}}$	Estimate	Interpretation
$e^{\hat{\beta}_1}$	1.011	odds ratio of obesity for a one-unit increase in age, holding race/ethnicity and gender constant
$e^{\hat{\beta}_2}$	1.426	odds ratio of obesity for Black non-hispanic vs. white non-Hispanic, holding age and gender constant
$e^{\hat{\beta}_3}$	1.234	odds ratio of obesity for women vs. men, holding age and race/ethnicity constant

Logistic regression

What is the estimate of the baseline **odds** of being obese from this model?

```
> exp(coef(logistic.m1)[1])

(Intercept)
  0.5001965
```

What is the estimate of the baseline **probability** of being obese from this model?

```
> exp(coef(logistic.m1)[1])/(1 + exp(coef(logistic.m1)[1]))

(Intercept)
  0.3334207
```

Logistic regression

Given what we've just found out about the risk of being obese in this population, how do we expect the **risk ratios** for our covariate effects to look like compared to the **odds ratios**?

Since the baseline probability is 0.33, we know we are in a region where the odds ratio will overestimate the risk ratio. Note that this doesn't mean the odds ratio is invalid; it just means that if we are tempted to interpret the odds ratio as a risk ratio, we should be careful!

Logistic regression

Let's also summarize the **risk ratios** for race/ethnicity and gender:

```
# Compute the risk ratios for race and gender at the mean of age
newdata <- data.frame(agectr=c(0,0,0,0),black=c(FALSE,TRUE,FALSE,TRUE),gender=c(1,1,2,2))
newdata$risks <- predict(logistic.m1, newdata, type="response")
print(newdata)

  agectr black female     risks
1      0 FALSE  FALSE 0.3334207
2      0  TRUE  FALSE 0.4163455
3      0 FALSE   TRUE 0.3815853
4      0  TRUE   TRUE 0.4680774
```

Variable	Among	Odds Ratio	Risk Ratio
black	female=0	1.426	1.249
black	female=1	1.426	1.227
gender	black=0	1.234	1.144
gender	black=1	1.234	1.124

GLMs and the canonical link

Structure of a Generalized Linear Model (GLM)

A GLM has three components:

1. **Random component:**

$$Y \sim \text{Exponential Family}(\theta, \phi)$$

2. **Systematic component:**

$$\eta = \mathbf{X}\beta$$

3. **Link function:**

$$g(\mu) = \eta \quad \text{where } \mu = \mathbb{E}(Y)$$

Exponential Family Representation

A distribution belongs to the exponential family if it can be written as:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

Key identity:

$$\mu = \mathbb{E}(Y) = b'(\theta)$$

Here, θ is called the **natural (or canonical) parameter**.

Exponential Family Components for Common Distributions

General form:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

Distribution	θ (natural param)	$a(\phi)$	$b(\theta)$	$c(y,\phi)$
Bernoulli(p)	$\log \frac{p}{1-p}$	1	$\log(1+e^\theta)$	0
Poisson(λ)	$\log \lambda$	1	e^θ	$-\log(y!)$
Normal(μ, known σ^2)	μ	σ^2	$\frac{1}{2}\theta^2$	$-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$

$$\mu = b'(\theta) \quad \Rightarrow \quad \text{canonical link: } g(\mu) = \theta$$

Definition of the Canonical Link

Canonical link:

$$g(\mu) = \theta$$

That is, the link function directly maps the mean μ to the natural parameter θ in the exponential family representation.

Examples of Canonical Links

Distribution	θ (natural parameter)	Canonical link
Normal	$\theta = \mu$	$g(\mu) = \mu$ (Identity)
Binomial	$\theta = \log \frac{\mu}{1-\mu}$	$g(\mu) = \log \frac{\mu}{1-\mu}$ (Logit)
Poisson	$\theta = \log(\mu)$	$g(\mu) = \log(\mu)$ (Log)

Why Use the Canonical Link?

Using the canonical link often provides:

- Simpler score equations:

$$\mathbf{X}^\top (Y - \mu) = 0$$

- Faster convergence during estimation
- Convenient sufficient statistics
- Cleaner theoretical results (e.g., orthogonality of parameters)

Summary

- GLMs connect mean response μ to predictors through a link.
- Exponential family distributions are parameterized by a natural parameter θ.
- The **canonical link** is the one that sets $g(\mu) = \theta$.
- It is “canonical” because it matches the intrinsic structure of the distribution.

BUT

Do we always need to use the canonical link? NO!

Log binomial regression

Log binomial regression

When $P(Y = 1)$ is not rare, the odds ratio will be larger (i.e. further from the null) than the risk ratio. Also, when adjusting for multiple covariates in a logistic regression model, the odds ratios will be constant across covariate strata but the risk ratios will not be. In general, when the outcome is not rare, we cannot interpret the odds ratios as relative risks.

Wacholder (1986) recommended fitting a **log binomial** model for the risk in a GLM framework. Here, we we model

$$\log[\pi(\mathbf{x})] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

This is a GLM with binomial error distribution but where the link is $\eta_i = \log[\pi_i]$.

Log binomial regression

```
logbin.m1 <- glm(obese ~ agectr + black + factor(gender),
                 data=d.obese,family=binomial(link="log"))
summary(logbin.m1)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.102464   0.043681 -25.239  < 2e-16 ***
agectr          0.006571   0.002226   2.952  0.00316 **
blackTRUE       0.214920   0.051337   4.186 2.83e-05 ***
femaleTRUE      0.137603   0.051107   2.692  0.00709 **
```

How do we interpret $e^{\hat{\beta}}$?

$e^{\hat{\beta}}$	Estimate	Interpretation
$e^{\hat{\beta}_0}$	0.332	risk (probability) of obesity at the mean of age, for white non-Hispanic men
$e^{\hat{\beta}_1}$	1.007	risk ratio of obesity for a one-unit increase in age, holding race/ethnicity and gender constant
$e^{\hat{\beta}_2}$	1.240	risk ratio of obesity for Black non-Hispanic vs. white non-Hispanic, holding age and gender constant
$e^{\hat{\beta}_3}$	1.148	risk ratio of obesity for women vs. men, holding age and race/ethnicity constant

Log Poisson regression

Log Poisson regression

Spiegelman and Hertzmark (2005) and numerous other authors note that the log binomial model is less stable than the logistic binomial model and may fail to converge. When this is the case, they recommend fitting a Poisson regression model and using robust variance estimates. On average, the modified Poisson estimates are valid but not fully efficient when compared with log-binomial maximum likelihood estimators (Spiegelman and Hertzmark, 2005).

```
poisson.m1 <- glm(obese ~ agectr + black + factor(gender),
                  data=d.obese,family=poisson(link="log"))
summary(poisson.m1)

robust.estimates <- cbind(coef(poisson.m1),
                          coef(poisson.m1) - 1.96*sqrt(diag(vcovHC(poisson.m1))),
                          coef(poisson.m1) + 1.96*sqrt(diag(vcovHC(poisson.m1))))
colnames(robust.estimates) <- c("Estimate","Lower 95% CI","Upper 95% CI")
print(exp(robust.estimates))
```

Log Poisson regression

	Estimate	Lower 95% CI	Upper 95% CI
(Intercept)	0.3342147	0.3065488	0.3643774
agectr	1.0067362	1.0023304	1.0111613
blackTRUE	1.2351684	1.1164364	1.3665275
femaleTRUE	1.1360580	1.0272361	1.2564081

Comparison

	Log binomial			Log Poisson		
	Estimate	2.5%	97.5%	Estimate	2.5%	97.5%
(Intercept)	0.3321	0.3039	0.3614	0.3342	0.3065	0.3644
agectr	1.0066	1.0022	1.0110	1.0067	1.0023	1.0112
blackTRUE	1.2398	1.1199	1.3702	1.2352	1.1164	1.3665
femaleTRUE	1.1475	1.0385	1.2688	1.1361	1.0272	1.2564

Which model do we prefer?

Model	Link	Interpretation of $\hat{\beta}$s
logistic binomial	$\log[\pi(x)/(1-\pi(x)]$	log odds ratios
log binomial	$\log[\pi(x)]$	log risk ratios
log Poisson	$\log[\pi(x)]$	log risk ratios

Some things to think about

- As noted above, the log binomial model is more unstable and may not converge, especially when estimates of $\hat{\pi}_i$ are near the boundaries.
- In the log Poisson models, the log link function could theoretically yield fitted values of π_i that are greater than one.
- Another peculiar thing about risks vs. odds: let's say that you are comparing two groups: one where the risk of obesity is 0.2 and the other where the risk is 0.4. The risk ratio for group 2 vs. group 1 is $RR = 0.4/0.2 = 2$.
- Now let's say that you are comparing the same two groups, but you are looking at the risk of **not** being obese. The risk in group 1 is 1-0.2=0.8. In the second group, the risk is 1-0.4=0.6. The risk ratio for for group 1 vs. group 2 is $RR = 0.8/0.6 = 1.33$.

Some things to think about

- Now let's compare the **odds** of being obese for group 2 vs. group 1. $OR = \frac{0.4/0.6}{0.2/0.8} = 2.667$.
- Now let's compare the **odds** of **not** being obese for group 1 vs. group 2. $OR = \frac{0.8/0.2}{0.6/0.4} = 2.667$.
- How do you decide which approach to take? Depends on how you want to interpret your estimates and what you think the underlying data generating process looks like.
- One option if you want to interpret risk ratios but think that the model is linear in the log odds is to fit the logistic binomial model and then generate predicted probabilities for covariate strata of interest. Then you can make risk comparisons based on the fitted probabilities.

References

1. Spiegelman D, Hertzmark H. (2005). Easy SAS calculations for risk or prevalence ratios and differences. *Am J Epidemiol* 162:199-200.
2. Wacholder S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol* 123: 174-84.