Lecture 17

Relative Risk, Odds Ratio, and Logistic Regression

STAT 8020 Statistical Methods II
October 20, 2020

Whitney Huang
Clemson University

Agenda

1. **Relative Risk and Odds Ratio**

2. **Logistic Regression**

Aspirin Use and Heart Attack

The table below is from a report on the relationship between aspirin use and heart attract by the Physicians' Health Study Research Group at Harvard Medical School (*New Engl. J. Med.* **318**: 262-264 ,1988).

	Heart Attack		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Here we want to know if the use of aspirin effectively reduces the heart attack rate. To do so we are going to introduce relative risk and odds ratio.

Relative Risk and Odds Ratio

The relative risk (RR) is defined to be the ratio

$$\mathrm{RR} = \frac{p_1}{p_2},$$

where p_i is the probability of “success” for the i^{th} group.

The odds ratio (θ) is defined as

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)},$$

where Ω is the odds for the i^{th} group.

Aspirin Use and Heart Attack Revisited: Inference for θ

	Heart Attack		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Point Estimation:

$$\hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}},$$

where n_{i1} is the number of successes and n_{i2} is the number of the failures of the i^{th} group. For this example, we have

$$\hat{\theta} = \frac{189 \times 10933}{10845 \times 104} = 1.83$$

$\Rightarrow$ the odds of heart attack for those taking placebo was 1.83 times the odds of those taking aspiring.

Inference for θ Cont'd

Interval Estimation:
Confidence interval is constructed in the nature log scale. We have the Wald confidence interval for $\log \theta$

$$\log \hat{\theta} + z_{\alpha/2}\hat{\sigma}_{\log \hat{\theta}},$$

where the estimated standard error for $\log \hat{\theta}$ is

$$\hat{\sigma}_{\log \hat{\theta}} = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)^{1/2}.$$

Exponentiating its end points (i.e., the lower limit and the upper limit) provides a confidence interval for θ.

Aspirin Use and Heart Attack: Confidence Interval for θ

Suppose we want to construct a 95% CI for θ:

- $\log \hat{\theta} = 0.6054$
- $\hat{\sigma}_{\log \hat{\theta}} = \left(\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933}\right)^{1/2} = 0.1228$
- Margin of error: $z_{0.025} \times \hat{\sigma}_{\log \hat{\theta}} = 1.96 \times 0.1228 = 0.2407$
- CI on the nature log scale:
  $[0.6054 - 0.2407, 0.6054 + 0.2407] = [0.3647, 0.8461]$
- CI on the original scale:
  $[\exp(0.3647), \exp(0.8461)] = [1.4401, 2.3305]$

Handedness vs. Gender Example Revisited

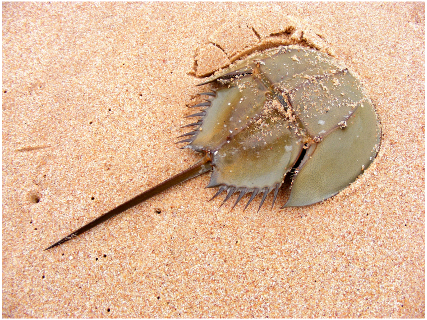
Here we'd like to use the table below to infer the male to female odds ratio of left-handedness

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Total	87	13	100

- Find the point estimate
- Construct a 95% confidence interval

Logistic Regression

A Motivating Example: Horseshoe Crab Malting [Brockmann, 1996, Agresti, 2013]



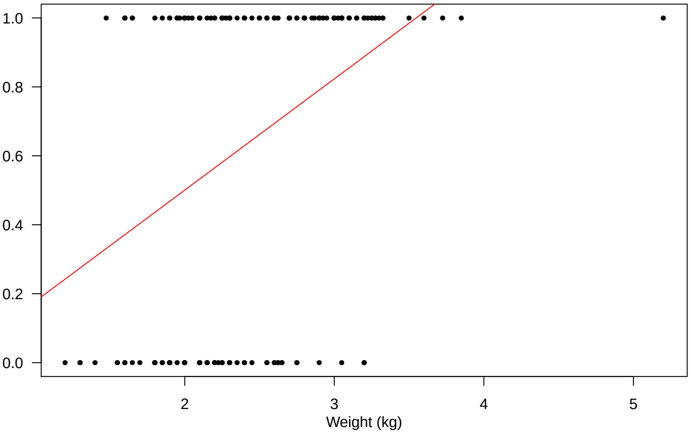
sat	y	weight	width
8	1	3.05	28.3
0	0	1.55	22.5
9	1	2.30	26.0
0	0	2.10	24.8
4	1	2.60	26.0
0	0	2.10	23.8
0	0	2.35	26.5
0	0	1.90	24.7
0	0	1.95	23.7
0	0	2.15	25.6

Source: `https://www.britannica.com/story/horseshoe-crab-a-key-player-in-ecology-medicine-and-more`

In the rest of today's lecture, we are going to use this data set to illustrate logistic regression. The response variable is y: whether there are males clustering around the female

Let’s Fit a Linear Regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{weight}$$



Fitting a linear regression to binary response is problematic. (Why?) We need a different statistical model to describe the data

Logistic Regression

Let $\mathrm{P}(Y = 1) = \pi \in [0, 1]$, and x be the predictor (`weight` in the previous example). In SLR we have

$$\pi(x) = \beta_0 + \beta_1 x,$$

which will lead to invalid estimate of π (i.e., > 1 or < 0).

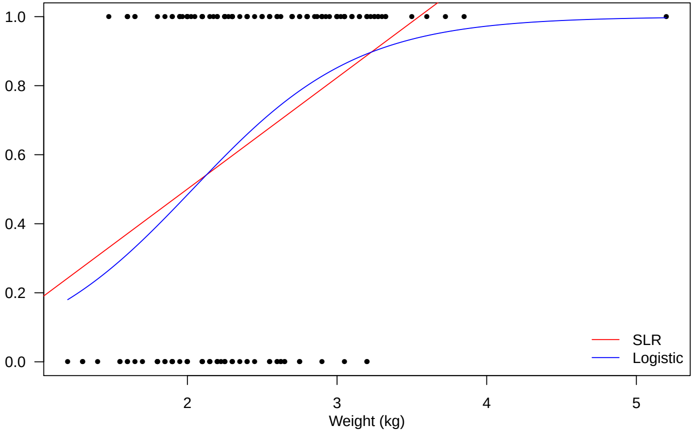
Logistic Regression

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x.$$

- $\log(\frac{\pi}{1-\pi})$: the log-odds or the logit
- $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \in (0, 1)$

Logistic Regression Fit

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}},\ \hat{\beta}_0 = -3.6947(0.8802),\ \hat{\beta}_1 = 1.8151(0.3767).$$



Properties

- Similar to SLR, Sign of β_1 indicates whether $\pi(x) \uparrow$ or $\downarrow$ as $x \uparrow$
- If $\beta_1 = 0$, then $\pi(x) = e^{\beta_0}/(1 + e^{\beta 0})$ is a constant w.r.t x (i.e., π does not depend on x)
- Curve can be approximated at fixed x by straight line to describe rate of change: $\frac{d\pi(x)}{dx} = \beta_1 \pi(x)(1 - \pi(x))$
- $\pi(-\beta_0/\beta_1) = 0.5$, and $1/\beta_1 \approx$ the distance of x values with $\pi(x) = 0.5$ and $\pi(x) = 0.75$ (or $\pi(x) = 0.25$)

Odds Ratio Interpretation

Recall $\log(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_1 x$, we have the odds

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0 + \beta_1 x).$$

If we increase x by 1 unit, the the odds becomes

$$\exp(\beta_0 + \beta_1(x+1)) = \exp(\beta_1) \times \exp(\beta_0 + \beta_1 x).$$

$$\Rightarrow \frac{\text{Odds at } x+1}{\text{Odds at } x} = \exp(\beta_1), \forall x$$

Example: In the horseshoe crab example, we have $\hat{\beta}_1 = 1.8151 \Rightarrow e^{1.8151} = 6.14 \Rightarrow$ Estimated odds of satellite multiply by 6.1 for 1 kg increase in weight.

Parameter Estimation

In logistic regression we use maximum likelihood estimation to estimate the parameters:

- **Statistical model**: $Y_i \sim$ Bernoulli$(\pi(x_i))$ where $\pi(x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$.
- **Likelihood function**: We can write the joint probability density of the data $\{x_i, y_i\}_{i=1}^n$ as

  $$\prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{(1-y_i)}.$$

  We treat this as a function of parameters (β_0, β_1) given data.
- **Maximum likelihood estimate**: The maximizer $\hat{\beta}_0, \hat{\beta}_1$ is the maximum likelihood estimate (MLE). This maximization can only be solved numerically.

Horseshoe Crab Logistic Regression Fit

```
> logitFit <- glm(y ~ weight, data = crab, family = "binomial")
> summary(logitFit)

Call:
glm(formula = y ~ weight, family = "binomial", data = crab)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1108  -1.0749   0.5426   0.9122   1.6285

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 195.74  on 171  degrees of freedom
AIC: 199.74

Number of Fisher Scoring iterations: 4
```

Inference: Confidence Interval

A 95% confidence interval of the parameter β_i is

$$\hat{\beta}_i \pm z_{0.025} \times \mathrm{SE}_{\hat{\beta}_i}, \quad i = 0, 1$$

Horseshoe Crab Example

A 95% (Wald) confidence interval of β_1 is

$$1.8151 \pm 1.96 \times 0.3767 = [1.077, 2.553]$$

Therefore a 95% CI of e^{β_1}, the multiplicative effect on odds of 1-unit increase in x, is

$$[e^{1.077}, e^{2.553}] = [2.94, 12.85]$$

Inference: Hypothesis Test

Null and Alternative Hypotheses:

$$H_0 : \beta_1 = 0 \Rightarrow Y \text{ is independent of } X \Rightarrow \pi(x) \text{ is a constant}$$
$$H_a : \beta_1 \neq 0$$

Test Statistics:

$$z_{obs} = \frac{\hat{\beta}_1}{\mathrm{SE}_{\hat{\beta}_1}} = \frac{1.8151}{0.3767} = 4.819.$$

P-value $= 1.45 \times 10^{-6}$

We have sufficient evidence that `weight` has positive effect on π, the probability of having satellite male horseshoe crabs