Analysis of binary outcomes and logistic regression

LAs BEST 2022

Eric S. Kawaguchi

Division of Biostatistics and Epidemiology
Department of Population and Public Health Sciences
University of Southern California

June 28, 2022

Keck School of Medicine of USC

Binary Outcomes

- $Y = 1$ if an event (i.e. outcome of interest) occurs and $Y = 0$ if it does not occur.
  - Might be inherently dichotomous: Dead/alive; Asthmatic/Non-asthmatic; Pass/Fail
  - Or can be created from a continuous variable: $Y = 1$ if SBP ≥ 140 mmHG and $Y = 0$ is SBP < 140 mmHG.
- Logistic regression is the most commonly-used regression model for binary outcomes.
- Goal of logistic regression is to identify if there is an association between X (set of predictors/covariates) and the Y (the binary outcome).

Keck School of Medicine of USC

Binary Outcomes and Proportions

- With $Y \in \{0, 1\}$, we are interested in the proportion of events;
- Population proportion is often defined as p;
- Sample proportion: $\hat{p}$:

$$\begin{aligned}\hat{p} &= \frac{\#\text{ of observations with } Y = 1}{\text{Total } \#\text{ of observations in our sample}} \\ &= \frac{\sum_{i=1}^{N} Y_i}{N} = \bar{Y}.\end{aligned}$$

  where N is the sample size and Y_i is the outcome for the ith individual.
- Second equality holds ONLY if $Y \in \{0, 1\}$, which is the typical coding.

Keck School of Medicine of USC

Binary Outcomes and Proportions

- The population proportion (p) is estimated by the sample proportion ($\hat{p}$).
- Assuming a random sample, the sample proportion is unbiased with estimated variance $\widehat{Var(\hat{p})} = \hat{p}(1-\hat{p})/N$.
- Given a "large enough N", a 95% confidence interval (CI) for $\hat{p}$ is given by:

$$\hat{p} \pm 1.96 \times \sqrt{\hat{p}(1-\hat{p})/N}.$$

- How large is "large enough"?
- One rule-of-thumb is $Np(1-p) > 5$.
- Exact methods exist when this condition is not met.

Keck School of Medicine of USC

Example: Children's Health Study (CHS) Data

- Outcome of interest: Asthma (Yes / No)
- Out of 996 children, 209 have reported asthma.
- $\hat{p} = 209/996 = 0.21$ (or 21%);
- A 95% confidence interval for $\hat{p}$ is (0.185, 0.235)

Keck School of Medicine of USC

Association Between Binary X and Binary Y

- We are interested in seeing if there is an association between biological sex (X) and asthma (Y).
- We can use a contingency table to look at associations between a categorical Y and categorical X variable.
  - When X and Y are binary we have a 2×2 table

Keck School of Medicine of USC

Association Between Binary X and Binary Y

- 2×2 tables have the following form:

	$Y = 0$	$Y = 1$	Total
$X = 0$	a	b	$a + b$
$X = 1$	c	d	$c + d$
Totals	$a + c$	$b + d$	N

- An association between X and Y means the proportion/frequency of the event $(Y = 1)$ "differs" between $X = 0$ and $X = 1$.

Keck School of Medicine of USC

Association Between Binary X and Binary Y

- CHS Data:

X/Y	No	Yes	Total
Female	418	99	517
Male	369	110	479
Totals	787	209	996

- Frequency of asthma among females: $99/517 \approx 0.19$

Keck School of Medicine of USC

Association Between Binary X and Binary Y

- CHS Data:

X/Y	No	Yes	Total
Female	418	99	517
Male	479	110	479
Totals	787	209	996

- Frequency of asthma among females: $99/517 \approx 0.19$
- Frequency of asthma among males: $110/479 \approx 0.23$

Keck School of Medicine of USC

Association Between Binary X and Binary Y

- CHS Data:

X/Y	No	Yes	Total
Female	418	99	517
Male	479	110	479
Totals	787	209	996

- Frequency of asthma among females: $99/517 \approx 0.19$
- Frequency of asthma among males: $110/479 \approx 0.23$
- **Question: Is there an association between asthma and biological sex?**

Keck School of Medicine of USC

The Odds and Odds Ratio

- Let $p = \Pr(Y = 1)$ and $1 - p = \Pr(Y = 0)$.
- The odds (of an event occurring) is defined as $p/(1 - p)$ and can be estimated by $\hat{p}/(1 - \hat{p})$.
- Some examples:
  - The odds of getting heads with a fair coin are $0.5/0.5 = 1$.
  - The odds of winning the lottery jackpot are 1 in 3.5 million.
  - "May the odds be ever in your favor" -Effie Trinket
- Events range between $\{0, 1\}$;
- Probabilities/frequencies/proportions range between $[0, 1]$;
- Odds range between $[0, \infty)$.

Keck School of Medicine of USC

The Odds and Odds Ratio

- The odds ratio is a ratio of two odds
- Odds ratio (for binary $X \in \{0, 1\}$):

$$OR = \frac{\text{Odds of } Y = 1 \text{ when } X = 1}{\text{Odds of } Y = 1 \text{ when } X = 0} = \frac{\frac{p_1}{(1-p_1)}}{\frac{p_0}{(1-p_0)}}$$

- The odds that Y will occur ($Y = 1$) when $X = 1$ compared to the odds that Y will occur when $X = 0$.

Keck School of Medicine of USC

The Odds and Odds Ratio

Strength of association:

- $OR = 1$: Odds of $Y = 1$ are the same for $X = 0$ and $X = 1$ (no association between X and Y);
- $OR > 1$: Odds of $Y = 1$ are larger for $X = 1$ than for $X = 0$ (positive association between X and Y);
- $OR < 1$: Odds of $Y = 1$ are smaller for $X = 1$ than for $X = 0$ (negative association between X and Y);

Keck School of Medicine of USC

The Odds and Odds Ratio

	$Y = 0$	$Y = 1$	Total
$X = 0$	a	b	$a + b$
$X = 1$	c	d	$c + d$
Totals	$a + c$	$b + d$	N

- The odds ratio is estimated by:

$$\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)} = \frac{ad}{bc}$$

Keck School of Medicine of USC

CHS Example

- The estimated odds of having asthma ($Y = 1$) in females ($X = 0$) is:

$$\hat{p}_0/(1 - \hat{p}_0) = 0.19/0.81 = 99/418 \approx 0.24.$$

- The estimated odds of having asthma ($Y = 1$) in males ($X = 1$) is:

$$\hat{p}_1/(1 - \hat{p}_1) = 0.23/0.77 = 110/369 \approx 0.30.$$

- The odds ratio (males relative to females) is estimated to be:

$$\widehat{OR} = 0.30/0.24 \approx 1.25$$

Keck School of Medicine of USC

CHS Example

- The estimated odds ratio is 1.25.
- Interpretation: The odds of males having asthma is 1.25 times the odds of females having asthma.
- Can construct confidence intervals and test if this odds ratio is significantly different from 1 ($H_0 : OR = 1$) using data from the 2×2 table.
  - χ^2 test (Large n)
  - Fisher's exact test (Small n)

Keck School of Medicine of USC

The Odds Ratio

- Interpretation of the odds ratio is not as straightforward as with a relative risk;
- If the disease is rare, in the general population we are studying, then the odds ratio approximates the relative risk;
- Both odds ratio or relative risk can be directly estimated from a cohort study
- Only the odds ratio can be directly estimated when cases are over represented (case-control study).

Keck School of Medicine of USC

Logistic Regression

Keck School of Medicine of USC

Simple Logistic Regression

- Oftentimes X may be categorical (with more than 2 categories) or continuous:
  - Hard to tabulate the data as with a 2×2 table
- Assume, for now, we only have one covariate/predictor X.
- For simple linear regression, we model $E(Y|X) = \beta_0 + \beta_1 X$, where $E(Y|X)$ is the mean/average value of Y given X.
  - β_0 is the mean value of Y when $X = 0$;
  - β_1 is the change in the mean value of Y for a one-unit change in X.

Keck School of Medicine of USC

Simple Logistic Regression

The simple logistic regression model:

$$\log(\pi(X)) = \log\left(\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}\right) = \beta_0 + \beta_1 X,$$

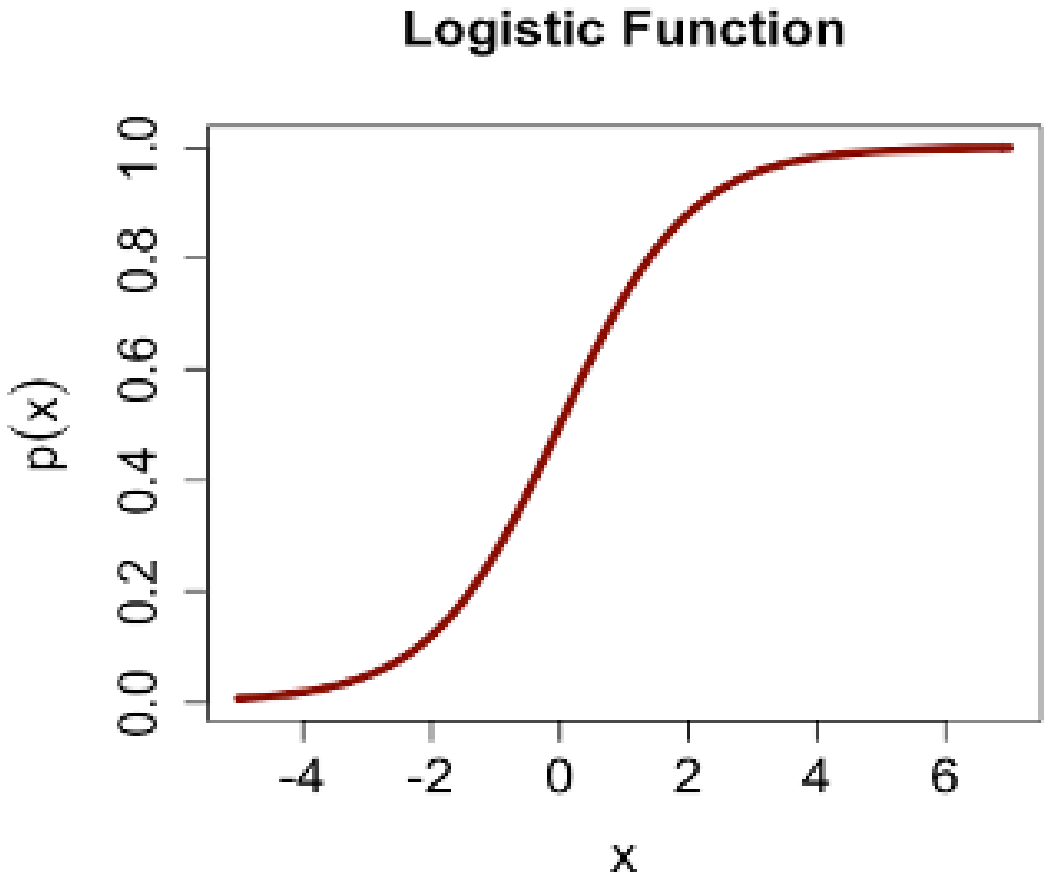
- $\pi(X)$ is the odds of $Y = 1$ given X.
- We assume linearity on the *logit* (i.e. log-odds) scale;
- It can be shown that:

$$\Pr(Y=1|X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}.$$

- $f(t) = \frac{\exp(t)}{1+\exp(t)}$ is called the *logistic* function.

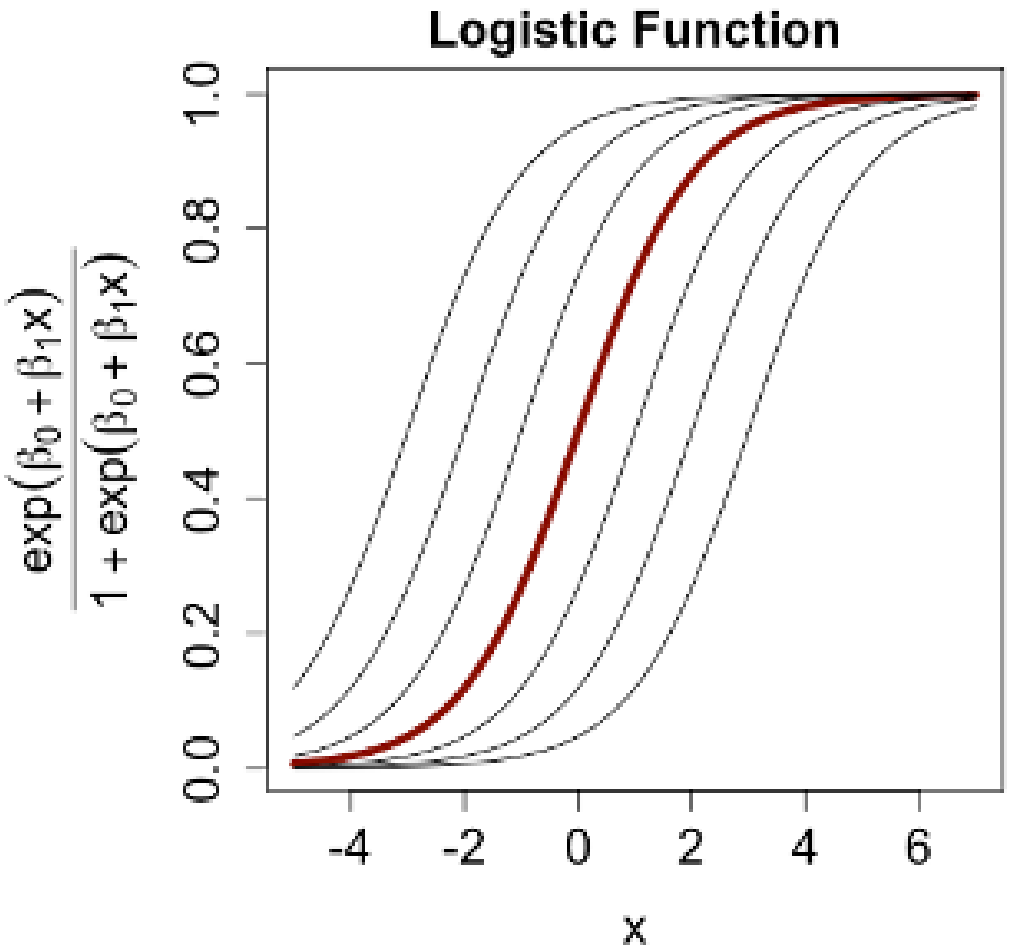
Keck School of Medicine of USC

Plot of the logistic function



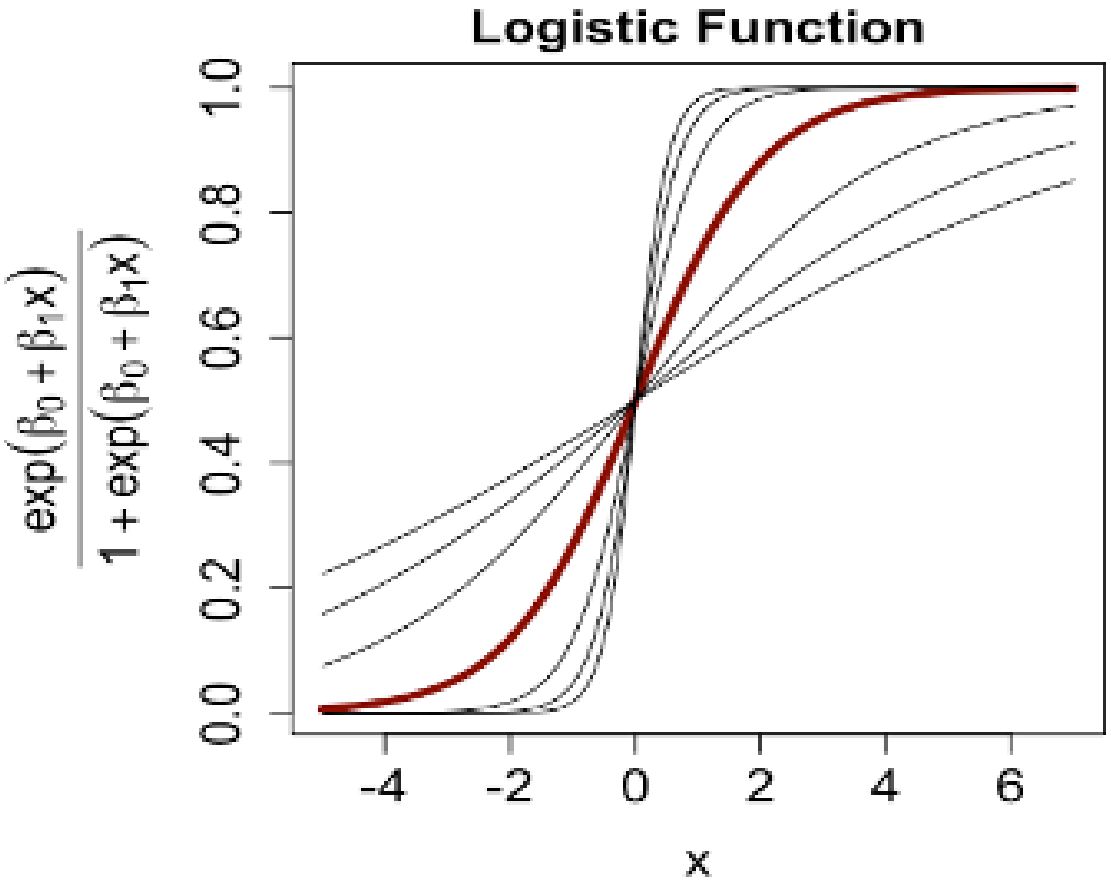
Keck School of Medicine of USC

Effect of β_0 on the logistic function



Keck School of Medicine of USC

Effect of β_1 on the logistic function



Keck School of Medicine of USC

Interpretation of β_1

- Log odds for $X = x$:

$$\log(\pi(X = x)) = \beta_0 + \beta_1 x$$

- Log odds for $X = x + 1$:

$$\log(\pi(X = x + 1)) = \beta_0 + \beta_1(x + 1) = \beta_0 + \beta_1 x + \beta_1$$

- Note

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

Keck School of Medicine of USC

Interpretation of β_1

- Exponentiating both sides:

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

- The LHS is an odds ratio!
- $\exp(\beta_1)$ corresponds to the odds ratio for a 1-unit increase in X;
- When X is binary, $\exp(\beta_1)$ is the odds ratio of the two groups.

Keck School of Medicine of USC

Interpretation of β_0

- Log odds for $X = 0$:

$$\log(\pi(X = 0)) = \beta_0 + \beta_1 \times 0 = \beta_0$$

- Exponentiating both sides:

$$\pi(X = 0) = \exp(\beta_0).$$

- Binary X: The odds that Y will occur for the reference group
  - If X is biological sex and $X = 0$ codes females, then it's the odds that Y will occur for females.
- Continuous X: The odds that Y will occur when $X = 0$.
  - May not have a meaningful interpretation for continuous X.
  - If X is age, then it's the odds that Y will occur at age 0.

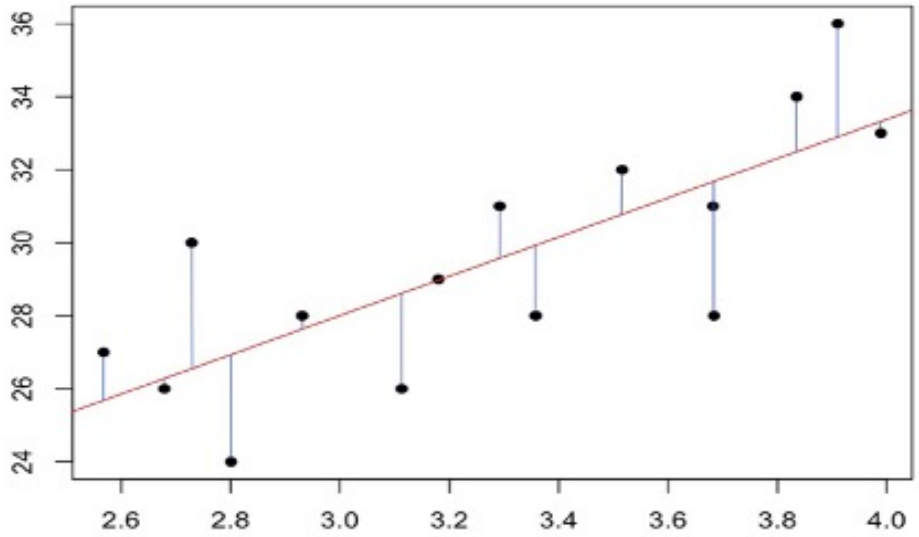
Keck School of Medicine of USC

Estimation of β_0, β_1

- In linear regression, the coefficients are estimated by least squares:

$$\arg\min_{\beta_0,\beta_1} \left\{ -\frac{1}{2} \| Y - (\beta_0 + \beta_1 X) \|_2^2 \right\}.$$

- The above formula is equivalent to finding the line that best fits the observed data: Minimizes the sum of the squared vertical distances from the observations to the line.



Keck School of Medicine of USC

Estimation of β_0, β_1

- In logistic regression we estimate β_0 and β_1 (and their corresponding standard errors) by minimizing an alternative measure of distance between the points *to the logistic curve*.
- This method is called *maximum likelihood estimation*.

Keck School of Medicine of USC

Hypothesis testing

- $\beta_1/\exp(\beta_1)$ quantifies the association between X and Y;
- We want to test if this association is statistically significant.
- The null hypothesis $H_0 : OR = 1$ or equivalently $H_0 : \beta_1 = 0$.
- One way to test this hypothesis is to compare the test statistic

$$Z = \frac{\hat{\beta}_1}{\widehat{se}(\hat{\beta}_1)}$$

  against the standard normal distribution $\mathcal{N}(0, 1)$.
- We reject H_0 if Z is large, or equivalently, if the p-value is small.

Keck School of Medicine of USC

Logistic Regression in R

- Logistic regression (logit model) can be fit in R used the `glm()` function.
- glm: Generalized linear model (will discuss later)

```
glm(y ~ x, family = "binomial", data = df)
```

Keck School of Medicine of USC

Binary X

```
lr_sex <- glm(asthma ~ sex, family="binomial", data=chs)
summary(lr_sex)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4404     0.1118 -12.886   <2e-16 ***
sexM          0.2300     0.1559   1.476     0.14
```

- $\hat{\beta}_1 = \exp(0.23) = 1.258$. p-value $= 0.14$, not statistically significant at the $\alpha = 0.05$ level;
- $\hat{\beta}_0 = \exp(-1.4404) = 0.236$

Keck School of Medicine of USC

CHS Example

From 2×2 table:

X/Y	No	Yes	Total
Female	418	99	517
Male	369	110	479
Totals	787	209	996

- $\text{OR} = ad/bc = (418 \times 110)/(99 \times 369) = 1.258$
- $\hat{\pi}_0 = (99/517)/(418/517) = 99/418 = 0.236$

Keck School of Medicine of USC

CHS Example

From 2×2 table:

X/Y	No	Yes	Total
Female	418	99	517
Male	369	110	479
Totals	787	209	996

- OR $= ad/bc = (418 \times 110)/(99 \times 369) = 1.258$
- $\hat{\pi}_0 = (99/517)/(418/517) = 99/418 = 0.236$
- Same results as the logistic regression model!

Keck School of Medicine of USC

Continuous X

```
lr_age <- glm(asthma ~ age, family = "binomial", data=chs)
summary(lr_age)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7521     1.9465   0.386    0.699
age          -0.1364     0.1278  -1.067    0.286
```

- $\hat{\beta}_1 = \exp(-0.136) = 0.87$. A 1-year increase in age is associated with a decrease in the odds of having asthma
- $\hat{\beta}_0 = \exp(0.752) = 2.12$

Keck School of Medicine of USC

Multivariable logistic regression

- For multivariable logistic logistic regression:

$$\log(\pi(X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p,$$

  where $X = (X_1, X_2, \ldots, X_p)$.
- $\hat{\beta}_1$: Log-odds ratio for Y associated with a one-unit change in X_1 while keeping all other variables (e.g. $X_2, \ldots, X_p$) fixed.
- $\hat{\beta}_2$: Log-odds ratio for Y associated with a one-unit change in X_2 while keeping all other variables (e.g. $X_1, X_3, \ldots, X_p$) fixed.
- And so on...

Keck School of Medicine of USC

Multivariable logistic regression

```
lr_sex_age = glm(asthma ~ sex + age,
                 family = "binomial", data = chs)
summary(lr_sex_age)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9373     1.9517   0.480    0.631
sexM          0.2491     0.1568   1.589    0.112
age          -0.1567     0.1285  -1.219    0.223
```

- $\hat{\beta}_{age} = \exp(-0.1567) = 0.85$.
- $\hat{\beta}_{age}$ is the estimated log-odds ratio for asthma associated with a change of one year increase in age but keeping biological sex category the same (M or F).

Keck School of Medicine of USC

In summary

- Logistic regression is the most commonly-used regression model for binary outcomes;
- Coefficient estimates are interpreted as log odds-ratios;
- Odds ratios can be estimated in all study designs (case-control, cross sectional, cohort, etc.);
- When the disease is rare in the general population, the odds ratio approximates the relative risk ratio.

Keck School of Medicine of USC

Some food for thought

- Logistic regression (in some form) can be used for polychotomous outcomes (nominal or ordinal).
- Logistic regression is a *generalized linear model* (GLM).
  - GLMs are typically used when trying to model non-normal, non-continuous data.
  - $g(x) = \log(x/(1-x))$ is known as a link function.
  - The log-odds interpretation comes from the use of this specific link function
  - Different outcomes will require different link functions (and different interpretations!)

Keck School of Medicine of USC

Aside: Why use the logit function?

- In "linear" regression,

$$E(Y|X) = \Pr(Y = 1|X) = \beta_0 + \beta_1 X = \eta$$

- We provide no restriction for β_0 and β_1.
- While η can be ANY real value, $E(Y|X) \in [0, 1]$.
- Can lead to nonsensical results (Predicted/estimated probabilities less than 0 or greater than 1).
- **Note: Just because models can run, doesn't mean the results are meaningful.**

Keck School of Medicine of USC

Aside: Why use the logit function?

- We want a function of $g(\cdot)$, so that $g[E(Y|X)] = g[\Pr(Y = 1|X)]$ can take on ANY real value.
- What about $g(x) = \log(\frac{x}{1-x})$?
  - $g(\cdot)$ the log-odds (logit function).

$$\log(\pi(X)) = \log\left(\frac{\Pr(Y = 1|X)}{\Pr(Y = 0|X)}\right) = \beta_0 + \beta_1 X$$

$$\Rightarrow \Pr(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}.$$

- Other functions for $g(\cdot)$ have been proposed.

Keck School of Medicine of USC