Logistic Regression

Odds ratios

Prof. Maria Tackett

STA 210

Click here for PDF of slides

Topics

- Use the odds ratio to compare the odds of two groups
- Interpret the coefficients of a logistic regression model with
  - a single categorical predictor
  - a single quantitative predictor
  - multiple predictors

Risk of coronary heart disease

This dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. We want to examine the relationship between various health characteristics and the risk of having heart disease.

`high_risk`:

- 1: High risk of having heart disease in next 10 years
- 0: Not high risk of having heart disease in next 10 years

`age`: Age at exam time (in years)

`education`: 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = College

High risk vs. education

	High Risk	Not High Risk
Some high school	323	1397
High school or GED	147	1106
Some college or vocational school	88	601
College	70	403

Compare the odds for two groups

	High Risk	Not High Risk
Some high school	323	1397
High school or GED	147	1106
Some college or vocational school	88	601
College	70	403

- We want to compare the risk of heart disease for those with a High School diploma/GED and those with a college degree.
- We'll use the **odds** to compare the two groups

Compare the odds for two groups

	High Risk	Not High Risk
Some high school	323	1397
High school or GED	147	1106
Some college or vocational school	88	601
College	70	403

$$\text{odds} = \frac{P(\text{success})}{P(\text{failure})} = \frac{\text{\# of successes}}{\text{\# of failures}}$$

Odds of being high risk for the **High school or GED** group

$$\frac{147}{1106} = 0.133$$

Odds of being high risk for the **College** group

$$\frac{70}{403} = 0.174$$

Based on this, we see those with a college degree had higher odds of being high risk of heart disease than those with a high school diploma or GED.

Odds ratio

	High Risk	Not High Risk
Some high school	323	1397
High school or GED	147	1106
Some college or vocational school	88	601
College	70	403

Let's summarize the relationship between the two groups. To do so, we'll use the **odds ratio (OR)**.

$$OR = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\omega_1}{\omega_2}$$

Odds ratio: College vs. High school or GED

	High Risk	Not High Risk
Some high school	323	1397
High school or GED	147	1106
Some college or vocational school	88	601
College	70	403

$$OR = \frac{\text{odds}_{College}}{\text{odds}_{HS}} = \frac{0.174}{0.133} = \mathbf{1.308}$$

The odds of being high risk of heart disease are 1.30 times higher for those with a college degree than those with a high school diploma or GED.

STA 210

Odds ratio: College vs. Some high school

	High Risk	Not High Risk
Some high school	323	1397
High school or GED	147	1106
Some college or vocational school	88	601
College	70	403

$$OR = \frac{\text{odds}_{College}}{\text{odds}_{SomeHS}} = \frac{70/403}{323/1397} = 0.751$$

The odds of being high risk of having heart disease for those with a college degree are 0.751 times the odds of being high risk for heart disease for those with some high school.

More natural interpretation

It's more natural to interpret the odds ratio with a statement with the odds ratio greater than 1.

The odds of being high risk for heart disease are 1.33 times higher for those with some high school than those with a college degree.

Logistic regression: categorical predictor

Recall: Education - 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = College

```
risk_model <- glm(high_risk ~ education,
                  data  = heart, family = "binomial")
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1.464	0.062	-23.719	0.000
education2	-0.554	0.107	-5.159	0.000
education3	-0.457	0.130	-3.520	0.000
education4	-0.286	0.143	-1.994	0.046

Interpreting `education4` - log-odds

term	estimate	std.error	statistic	p.value
(Intercept)	-1.464	0.062	-23.719	0.000
education2	-0.554	0.107	-5.159	0.000
education3	-0.457	0.130	-3.520	0.000
education4	-0.286	0.143	-1.994	0.046

The **log-odds** of being high risk of heart disease are expected to be 0.286 less for those with a college degree compared to those with some high school (the baseline group).

Interpreting `education4` - odds

term	estimate	std.error	statistic	p.value
(Intercept)	-1.464	0.062	-23.719	0.000
education2	-0.554	0.107	-5.159	0.000
education3	-0.457	0.130	-3.520	0.000
education4	-0.286	0.143	-1.994	0.046

The **odds** of being high risk of heart disease for those with a college degree are expected to be 0.751 (exp(-0.286)) **times** the odds for those with some high school.

Coefficients + odds ratios

The model coefficient, -0.286, is the expected change in the log-odds when going from the *Some high school* group to the *College* group.

Therefore, $\exp\{-0.286\}$ = 0.751 is the expected change in the **odds** when going from the *Some high school* group to the *College* group.

$$OR = \exp\{\hat{\beta}_j\} = e^{\hat{\beta}_j}$$

Logistic regression: quantitative predictor

term	estimate	std.error	statistic	p.value
(Intercept)	-5.619	0.288	-19.498	0
age	0.076	0.005	14.174	0

Interpreting `age`: log-odds

term	estimate	std.error	statistic	p.value
(Intercept)	-5.619	0.288	-19.498	0
age	0.076	0.005	14.174	0

For each additional year in age, the log-odds of being high risk of heart disease are expected to increase by 0.076.

Interpretating `age`: odds

term	estimate	std.error	statistic	p.value
(Intercept)	-5.619	0.288	-19.498	0
age	0.076	0.005	14.174	0

For each additional year in age, the odds of being high risk of heart disease are expected to multiply by a factor of 1.08 (exp(0.076)).

Alternate interpretation

For each additional year in age, the odds of being high risk for heart disease are expected to increase by 8%.

Logistic regression: multiple predictors

```r
risk_model_3 <- glm(high_risk ~ education + age,
                    data = heart, family = "binomial")
```

term	estimate	std.error	statistic	p.value
(Intercept)	-5.385	0.308	-17.507	0.000
education2	-0.242	0.112	-2.162	0.031
education3	-0.235	0.134	-1.761	0.078
education4	-0.020	0.148	-0.136	0.892
age	0.073	0.005	13.385	0.000

Interpretation in terms of the log-odds

term	estimate	std.error	statistic	p.value
(Intercept)	-5.385	0.308	-17.507	0.000
education2	-0.242	0.112	-2.162	0.031
education3	-0.235	0.134	-1.761	0.078
education4	-0.020	0.148	-0.136	0.892
age	0.073	0.005	13.385	0.000

`education4`: The **log-odds** of being high risk of heart disease are expected to be 0.020 less for those with a college degree compared to those with some high school, **holding age constant.**

Interpretation in terms of the log-odds

term	estimate	std.error	statistic	p.value
(Intercept)	-5.385	0.308	-17.507	0.000
education2	-0.242	0.112	-2.162	0.031
education3	-0.235	0.134	-1.761	0.078
education4	-0.020	0.148	-0.136	0.892
age	0.073	0.005	13.385	0.000

age: For each additional year in age, the log-odds of being high risk of heart disease are expected to increase by 0.073, **holding education level constant.**

Interpretation in terms of the odds

term	estimate	std.error	statistic	p.value
(Intercept)	-5.385	0.308	-17.507	0.000
education2	-0.242	0.112	-2.162	0.031
education3	-0.235	0.134	-1.761	0.078
education4	-0.020	0.148	-0.136	0.892
age	0.073	0.005	13.385	0.000

`education4`: The **odds** of being high risk of heart disease for those with a college degree are expected to be 0.98 (exp(-0.020)) **times** the odds for those with some high school, **holding age constant**.

Interpretation in terms of the odds

term	estimate	std.error	statistic	p.value
(Intercept)	-5.385	0.308	-17.507	0.000
education2	-0.242	0.112	-2.162	0.031
education3	-0.235	0.134	-1.761	0.078
education4	-0.020	0.148	-0.136	0.892
age	0.073	0.005	13.385	0.000

age: For each additional year in age, the odds being high risk of heart disease are expected to multiply by a factor of 1.08 (exp(0.073)), **holding education level constant**.

Recap

- Use the odds ratio to compare the odds of two groups
- Interpret the coefficients of a logistic regression model with
  - a single categorical predictor
  - a single quantitative predictor
  - multiple predictors