Odds and Odds Ratio

Logistic Regression

Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 7.

Logistic Regression Model for Binary Y

$$Y = 0 \text{ or } 1$$

$$Z = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

What is the meaning of b_1, b_2, … b_m?

Odds of Event A

$$Odds(A) \equiv \frac{P(A)}{1 - P(A)}$$

Typically expressed as two numbers: Integer numerator and Integer denominator.

P(A) can be any probability function.

Example: Odds of Heart Attack

- Event A: Heart Attack
- If P(A) = 0.25, what is the Odds(A)?

Odds(A) = 0.25/(1-0.25) = 1/3
Odds of A is 1 to 3.

- If P(A) = 0.75, what is the Odds(A)?

Odds(A) = 0.75/(1-0.75) = 3/1
Odds of A is 3 to 1.

Odds of Event A if P(A) is the logistic function

Let A be the event Y = 1.

$$P(A) = P(Y = 1) = \frac{1}{1 + e^{-z}}$$

$$Odds(A) = Odds(Y = 1) \equiv \frac{P(Y = 1)}{1 - P(Y = 1)} = \frac{1}{1 + e^{-z}} \div \frac{e^{-z}}{1 + e^{-z}} = e^{z}$$

i.e. Odds of Y = 1 is exponentiation of the linear equation Z

How to isolate the model coefficient from e^z?

$$Odds(Y=1) = e^z = e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m}$$

- The model coefficients b_1, b_2, …b_m are inside the power of e.
- To isolate each of them, recall the formula:

$$\frac{a^m}{a^n} = a^{m-n}$$

Use the denominator with the same base to cancel all the terms that you don't want from m.

Odds Ratio for Continuous X_k

$$z = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m$$

$$OR_{X_k}(Y = 1) \equiv \frac{Odds_{X_k + 1}(Y = 1)}{Odds_{X_k}(Y = 1)} = e^{b_k}$$

For every 1 unit increase in x_k, the odds of Y = 1 multiply by e^{b_k}

If OR > 1, then increasing X_k will increase the odds of Y = 1, and vice versa.

Odds Ratio for Categorical X_k

Identify the baseline reference level $e.g.\ X_k = A$

$$OR_{X_k}(Y = 1) \equiv \frac{Odds_{X_k=B}(Y = 1)}{Odds_{X_k=A}(Y = 1)} = e^{b_k}$$

If x_k changes level from A to B, the odds of Y = 1 multiply by e^{b_k}

If OR > 1, then if X_k change from A to B, it will increase the odds of Y = 1, and vice versa.

Pass exam example: What is the meaning of the model coefficient 1.5046?

$$z = -4.0777 + 1.5046(Hours)$$

Hours is a continuous variable.

$$OR_{Hours}(Y=1) \equiv \frac{Odds_{Hours+1}(Y=1)}{Odds_{Hours}(Y=1)} = e^{1.5046} \approx 4.5$$

Studying for one additional hour will increase the odds of passing the exam by a factor of 4.5.

What if OR_X $(Y = 1) = 1$?

- X does not affect Odds of Y =1.
- 1 is the benchmark number to watch out for in any OR.
  - OR is just a fraction.
- OR > 1 means Odds of Y = 1 will increase if X changes in a specific direction.
- OR < 1 means Odds of Y = 1 will decrease if X changes in a specific direction.
- What if OR = 0.999876?
  - Considered as OR = 1?
  - Use either the p-value of X or the OR confidence interval to decide
  - Check if OR 95% confidence interval includes 1 or not.

Get OR and OR CI from R

```
> OR <- exp(coef(pass.m1))
> OR
(Intercept)       Hours
 0.01694617  4.50255687
```

e^{b_1}

```
> OR.CI <- exp(confint(pass.m1))
Waiting for profiling to be done...
> OR.CI
                   2.5 %     97.5 %
(Intercept) 0.0001868263  0.2812875
Hours       1.6978380343 23.2228735
```

95% CI excludes 1. Thus, Hours is statistically significant and increasing Hours will increase the odds of passing exam.

What's the difference between Odds vs Odds Ratios?

Odds

- Defined for the entire linear equation Z.
- e^z
- Is a function as z is a function.
- Measures the "chance" of Y = 1 using all the entire attributes $X_1, X_2, \ldots, X_m$.

Odds Ratio

- Defined for each model coefficient b_k
- e^b
- Is a number as b_k is a number.
- Measures the contribution of one attribute X_k to the Odds of Y = 1.

Next: Logistic Regression for Multi-categorical Y

- What if Y has more than 2 categories?
- Mathematical notation can be simplified and hidden for Binary Y.
- Suffice to consider the case where Y has 3 categories.