Logistic Regression for Y with more than 2 Categories

Logistic Regression

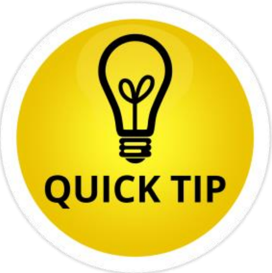
Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 7.

Y has 3 or More Possible Categories

- A/B/C/D/E
- Pass/Fail/Inconclusive
- 0/1/2
- Ang Mo Kio, Bedok, Clementi, ….
- Etc…

The baseline Reference Level for Y

- Y = 0 serves as the baseline.
- Dummy Variable Concept for Categorical X applies to Y too.



- Even if Y = A, B, C, [i.e. not 0, 1, 2], we can reduce mental effort by mapping Y = 0 to be the actual baseline reference level e.g. Y = A.
- Thus, all the formulas can still apply without any change to notation.
- i.e. it does not matter whether one label categorical Y as A, B, C, D or 0, 1, 2, 3. They are just labels for different categories of Y.

Logistic Regression Model for Binary Y

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

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

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

Easier to extend to multicategory Y

Two Outcomes Y vs Three Outcomes Y

$$Y = 0 \; or \; 1$$

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

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

Denominator to include all Zs.

$$P(Y = 0) = 1 \; - P(Y = 1)$$

$$Y = 0, 1 \; or \; 2$$

$$z_1 = b_{1,0} + b_{1,1} X_1 + b_{1,2} X_2 + \cdots + b_{1,m} X_m$$

$$z_2 = b_{2,0} + b_{2,1} X_1 + b_{2,2} X_2 + \cdots + b_{2,m} X_m$$

$$P(Y = 1) = \frac{e^{z_1}}{1 + e^{z_1} + e^{z_2}}$$

$$P(Y = 2) = \frac{e^{z_2}}{1 + e^{z_1} + e^{z_2}}$$

$$P(Y = 0) = 1 \; - P(Y = 1) - P(Y = 2)$$

Two Outcomes Y vs Three Outcomes Y

$$Y = 0 \; or \; 1$$

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

$$Odds(Y=1) \equiv \frac{P(Y=1)}{1 - P(Y=1)} \equiv \frac{P(Y=1)}{P(Y=0)} = e^z$$

$$Y = 0, 1 \; or \; 2$$

$$z_1 = b_{1,0} + b_{1,1} X_1 + b_{1,2} X_2 + \cdots + b_{1,m} X_m$$

$$z_2 = b_{2,0} + b_{2,1} X_1 + b_{2,2} X_2 + \cdots + b_{2,m} X_m$$

$$Odds(Y=1) \equiv \frac{P(Y=1)}{P(Y=0)} = e^{z_1}$$

$$Odds(Y=2) \equiv \frac{P(Y=2)}{P(Y=0)} = e^{z_2}$$

Summary for Multicategory Y

Y = 0 or 1	Y = 0, 1, or 2	Y = 0, 1, 2, or 3	Y = 0, 1, 2,…, k – 1
Y has 2 categories	Y has 3 categories	Y has 4 categories	Y has k categories

Summary for Multicategory Y

Y = 0 or 1	Y = 0, 1, or 2	Y = 0, 1, 2, or 3	Y = 0, 1, 2,…, k – 1
Y has 2 categories	Y has 3 categories	Y has 4 categories	Y has k categories
1 linear equation z	2 linear equations z_1 and z_2.	3 linear equations z_1 , z_2 and z_3.	k – 1 linear equations z_1 , z_2 and z_{k-1}.

Summary for Multicategory Y

Y = 0 or 1	Y = 0, 1, or 2	Y = 0, 1, 2, or 3	Y = 0, 1, 2,…, k – 1
Y has 2 categories	Y has 3 categories	Y has 4 categories	Y has k categories
1 linear equation z	2 linear equations z_1 and z_2.	3 linear equations z_1 , z_2 and z_3.	k – 1 linear equations z_1 , z_2 and z_{k-1}.
1 Odds e^z	2 Odds e^{z1} and e^{z2}	3 Odds e^{z1} , e^{z2} and e^{z3}	k – 1 Odds e^{z1} , e^{z2} and e^{z_k-1}

Summary for Multicategory Y

Y = 0 or 1	Y = 0, 1, or 2	Y = 0, 1, 2, or 3	Y = 0, 1, 2,…, k – 1
Y has 2 categories	Y has 3 categories	Y has 4 categories	Y has k categories
1 linear equation z	2 linear equations z_1 and z_2.	3 linear equations z_1 , z_2 and z_3.	k – 1 linear equations z_1 , z_2 and z_{k-1}.
1 Odds e^z	2 Odds e^{z1} and e^{z2}	3 Odds e^{z1} , e^{z2} and e^{z3}	k – 1 Odds e^{z1} , e^{z2} and e^{z_k-1}
2 probabilities	3 probabilities	4 probabilities	k probabilities

Computing Exercise in Ex 7.1

- Dataset: ratings.csv
- Y is Service Rating:
  - Bad
  - Neutral
  - Good
- Try to complete this exercise before class.

Summary

- Logistic Regression:
  - Predicting categorical Y
  - From linear equation that combines all Xs to Probability of Y, via logistic function.
  - From probability of Y to model predicted category of Y, via threshold.
  - From predicted categories to Confusion Matrix, by comparing model predicted categories vs actual categories of Y.
  - Binary Y vs Multi-categorical Y.
  - Main weakness of Logistic Regression – Perfect Separation.
    - As number of X increases, risk of perfect separation increases.
    - Do Ex 7.1.
- Odds Ratio:
  - Interpretation of each model coefficients.
  - Identify and Quantify Risk Factors.
    - Freitas et. al. (2012) Factors influencing hospital high length of stay outliers, BMC Health Services Research vol 12:265.