Overview of Logistic Regression

Logistic Regression

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

Objective

1. Apply an Analytics Model that can predict a Categorical Outcome Variable Y.
2. How to Identify & Quantify High Risk Factors.

Content

- Logistic Regression Model for Y with 2 outcomes (Binomial)
- Odds
- Odds Ratios and Odds Ratio Confidence Interval
- Logistic Regression Model for Y with 3 or more outcomes (Multinomial)

Next: The problem with Categorical Y

- Understand the problem.
- Understand why Logistic Regression is a Solution to that problem.

The Problem with Categorical Y

Logistic Regression

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

The most important idea in Linear Regression

- Linear Regression applies only for continuous Y.
  Example: $Y = 5 + 2X_1 - 3X_2$
- Most important idea is the possibility to use other information/predictors (i.e. Xs) to estimate Y:
  - Example: Estimate Price of a Flat (the Y variable).
  - X_1: Location
  - X_2: Level of the Flat
  - X_3: Within 5 mins walking distance to MRT?
  - X_4: Size of the Flat
  - X_5: Years remaining (max 99 years lease).
  - Etc…
- Use other predictors to predict certain disease, stock price, fraud, etc…
- Xs are unrestricted.

What if Y is categorical?

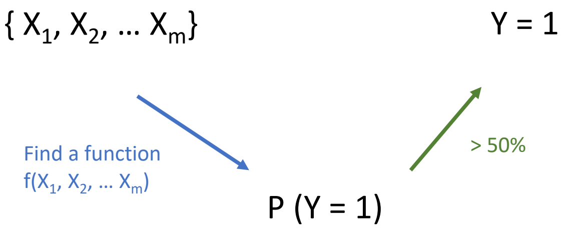
- Simplest Scenario: Y has only 2 Categorical levels (i.e. Binary Y)
  - 0 or 1
  - A or B
  - Yes or No
  - Pass or Fail
- Note: Logistic Regression can handle multi-categorical Y.
- Xs are unrestricted
  - Can be continuous or categorical.

The problem with Categorical Y

- How can we get a model to predict categorical Y, given a list of unrestricted Xs?
- Can $Y = 5 + 2X_1 - 3X_2$?
- In Linear Regression, Y is continuous implies Y can take any value within a certain range.
- Y is categorical means Y can only be A, B, C,….categorical levels.
- It is extremely difficult to find a linear equation involving unrestricted Xs that results in a categorical value for Y.
  - $b_0 + b_1X_1 + b_2X_2$ = category A ?

Reduce to a Simpler Problem: Find P(Y = 1)

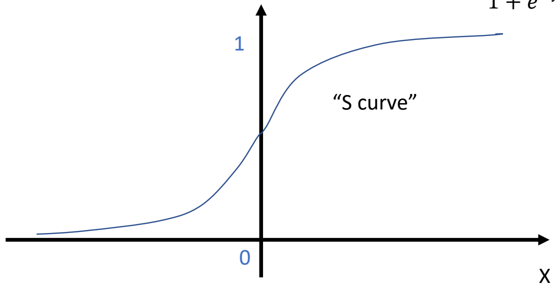
- Y = 0 or 1, and thus, has only two possible categorical values.
- P(Y = 1) has infinitely possible values within 0 to 1. i.e. continuous.
- P(Y = 0) can then be derived from P(Y = 1).



Logistic Function is suitable

- Unrestricted X
- Output is always between 0 to 1.
- Logistic Function f(x) can serve as Probability function P(Y = 1)

$$f(x) = \frac{1}{1 + e^{-x}}$$

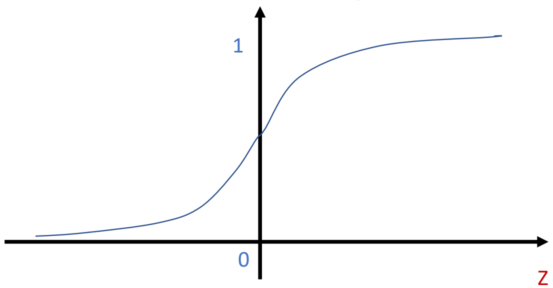


Logistic Function can admit many Xs.

- Unrestricted set of Xs by using Z to combine all Xs into one value.

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

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



Logistic Regression Model for Binary Y

Categorical Y = 0 or 1

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

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

Next: Numerical Example of Logistic Regression

- Strengthen your understanding by
- Applying the Logistic function
- On Simple Data from Wikipedia: passexam.csv, and
- Interpret the results

Logistic Regression for Y with 2 Categories

Logistic Regression

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

Logistic Regression Model for Binary Y

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

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

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

Example: Predicting Pass/Fail Exam

Source: Wikipedia

- A group of 20 students sat for an exam. Question: How does the number of hours spent studying affect the outcome of the exam (Pass/Fail)?
- The dataset (passexam.csv) shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

	A	B
1	Hours	Outcome
2	0.5	0
3	0.75	0
4	1	0
5	1.25	0
6	1.5	0
7	1.75	0
8	1.75	1
9	2	0
10	2.25	1
11	2.5	0
12	2.75	1
13	3	0
14	3.25	1
15	3.5	0
16	4	1
17	4.25	1
18	4.5	1
19	4.75	1
20	5	1
21	5.5	1

Base R: glm() function with family = binomial

Rscript: passexam.R

```r
library(data.table)

setwd('D:/Dropbox/Datasets/ADA1/7_Logistic_Reg')

passexam.dt <- fread("passexam.csv")
passexam.dt$Outcome <- factor(passexam.dt$Outcome)

summary(passexam.dt)

pass.m1 <- glm(Outcome ~ Hours , family = binomial, data = passexam.dt)

summary(pass.m1)
```

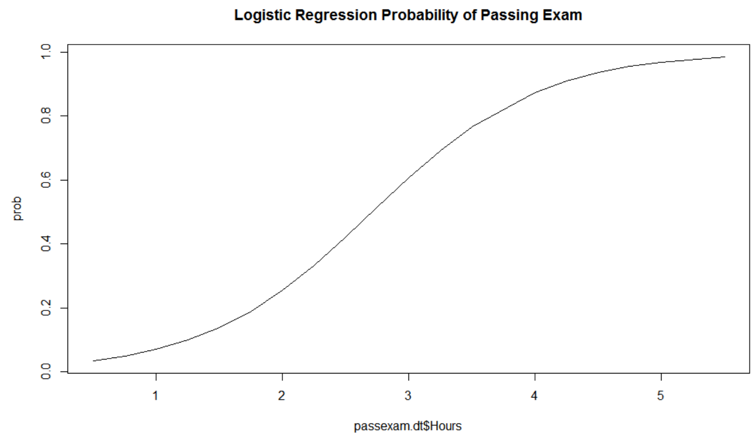

Results from summary() function

```
> summary(pass.m1)

Call:
glm(formula = Pass ~ Hours, family = binomial, data = passexam.dt
)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.70557  -0.57357  -0.04654   0.45470   1.82008

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0777     1.7610  -2.316   0.0206 *
Hours         1.5046     0.6287   2.393   0.0167 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hours is statistically significant factor.

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

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




```r
# Output the probability from the logistic function for all cases in the data.
prob <- predict(pass.m1, type = 'response')
# See the S curve
plot(x = passexam.dt$Hours, y = prob, type = "l", main = 'Logistic Regression Probability of Passing Exam')
```

Outputs the logistic function P(Y = 1)

Get model predicted Y (i.e. y.hat) by comparing P(Y = 1) against threshold.
Then get Confusion Matrix by comparing y.hat vs actual Y.

```r
# Set the threshold for predicting Y = 1 based on probability.
threshold <- 0.5

# If probability > threshold, then predict Y = 1, else predict Y = 0.
y.hat <- ifelse(prob > threshold, 1, 0)

# Create a confusion matrix with actuals on rows and predictions on columns.
table(passexam.dt$Outcome, y.hat, deparse.level = 2)

# Overall Accuracy
mean(y.hat == passexam.dt$Outcome)
```

```
> table(passexam.dt$Outcome, y.hat, deparse.level = 2)
                   y.hat
passexam.dt$Outcome 0 1
                  0 8 2
                  1 2 8
> # Overall Accuracy
> mean(y.hat == passexam.dt$Outcome)
[1] 0.8
```

Display Row name and Column name

What is the meaning of the model coefficient?

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

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

Next: Odds and Odds Ratio.

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 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.

Logistic Regression for Y with more than 2 Categories

Logistic Regression

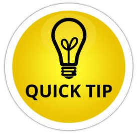
Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 7.

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 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

$Y = 0 \text{ or } 1$	$Y = 0, 1 \text{ or } 2$
$z = b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m$	$z_1 = b_{1,0} + b_{1,1}X_1 + b_{1,2}X_2 + \cdots + b_{1,m}X_m$ $z_2 = b_{2,0} + b_{2,1}X_1 + b_{2,2}X_2 + \cdots + b_{2,m}X_m$
$P(Y = 1) = \frac{1}{1 + e^{-z}} = \frac{e^{z}}{1 + e^{z}}$	$P(Y = 1) = \frac{e^{z_1}}{1 + e^{z_1} + e^{z_2}}$ $P(Y = 2) = \frac{e^{z_2}}{1 + e^{z_1} + e^{z_2}}$
Denominator to include all Zs.	
$P(Y = 0) = 1 - P(Y = 1)$	$P(Y = 0) = 1 - P(Y = 1) - P(Y = 2)$

Two Outcomes Y vs Three Outcomes Y

$Y = 0 \text{ or } 1$	$Y = 0, 1 \text{ or } 2$
$z = b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m$	$z_1 = b_{1,0} + b_{1,1}X_1 + b_{1,2}X_2 + \cdots + b_{1,m}X_m$ $z_2 = b_{2,0} + b_{2,1}X_1 + b_{2,2}X_2 + \cdots + b_{2,m}X_m$
$Odds(Y = 1) \equiv \frac{P(Y = 1)}{1 - P(Y = 1)} \equiv \frac{P(Y = 1)}{P(Y = 0)} = e^{z}$	$Odds(Y = 1) \equiv \frac{P(Y = 1)}{P(Y = 0)} = e^{z_1}$ $Odds(Y = 2) \equiv \frac{P(Y = 2)}{P(Y = 0)} = e^{z_2}$

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.