# Tut 3: Logistic Regression

June 2024

LR with numeric inputs $\boldsymbol{x} = (x_1, \cdots, x_p)$ only:

$$\mathbb{P}(Y = 1|\boldsymbol{x}) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p))}$$

LR with a $K$-level ($K \geq 2$) categorical input / qualitative predictor $X_i$:

$$\mathbb{P}(Y = 1|\boldsymbol{X}) = \frac{1}{1 + \exp(-(\beta_0 + \cdots + \beta_i^{(2)} x_i.\text{level2} + \cdots + \beta_i^{(K)} x_i.\text{level}K + \cdots))}$$

where $x_i.\text{level}k = \begin{cases} 1, & x_i = \text{level } k, \\ 0, & \text{otherwise} \end{cases}$, $k = 2, \cdots, K$.

$$Odds = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} = \frac{\mathbb{P}(Y=1)}{1-\mathbb{P}(Y=1)} = \frac{\frac{\exp(\dots)}{\exp(\dots)+1}}{1-\frac{\exp(\dots)}{\exp(\dots)+1}}$$

$$= \frac{\exp(\dots)}{\exp(\dots)+1-\exp(\dots)} = \exp(\dots) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p).$$

Let $k = 2, \dots, K$. Odds Ratio for numeric value:

$$OR = \frac{Odds(Y=1|X_i=b)}{Odds(Y=1|X_i=a)} = \frac{\exp(\cdots + \beta_i \cdot b + \dots)}{\exp(\cdots + \beta_i \cdot a + \dots)} = \exp(\beta_i(b-a)).$$

Odds Ratio for "one-hot-encoded" categorical value:

$$OR = \frac{Odds(Y=1|x_i.\text{level}k=1)}{Odds(Y=1|x.\text{level}k=0)} = \frac{\exp(\cdots + \beta_i^{(k)} \cdot 1 + \dots)}{\exp(\cdots + \beta_i^{(k)} \cdot 0 + \dots)} = \exp(\beta_i^{(k)}).$$

1. (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will default? [Answer: 27%]

   *Solution.* $\frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = 0.37 \Rightarrow \mathbb{P}(Y=1|X) = \frac{0.37}{1+0.37} = 0.270073$
   $\therefore$ fraction/probability = 27% □

   (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default? [Answer: 19%]

   *Solution.* odds $= \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \frac{0.16}{1-0.16} = 0.190048$
   $\therefore$ odds = 19%. □

2. The following table shows the results from logistic regression for ISLR **Weekly** dataset, which contains weekly returns of stock market (1 for up; 0 for down), based on predictors `Lag1` until `Lag5` and `Volume`.

| | Coefficient | Std. error | $Z$-statistic | $P$-value |
|---|---|---|---|---|
| Intercept | 0.2669 | 0.0859 | 3.11 | 0.0019 |
| Lag1 | -0.0413 | 0.0264 | -1.56 | 0.1181 |
| Lag2 | 0.0584 | 0.0269 | 2.18 | 0.0296 |
| Lag3 | -0.0161 | 0.0267 | -0.60 | 0.5469 |
| Lag4 | -0.0278 | 0.0265 | -1.05 | 0.2937 |
| Lag5 | -0.0145 | 0.0264 | -0.55 | 0.5833 |
| Volume | -0.0227 | 0.0369 | -0.62 | 0.5377 |

(a) Discuss how each predictor affects the weekly returns of stock market.

*Solution.* The predictors `Lag1`, `Lag3`, `Lag4`, `Lag5` and volume (with a **negative** coefficient, $\beta_i$) have a negative coefficients. When either one increases, the probability for weekly returns of stock market to increase is **lower**.

`Lag2` has a **positive** coefficient of 0.0584. Hence, when `Lag2` increases, the probability for weekly returns of stock market to increase is **higher**.

Mathematical derivation for the case $\beta_i < 0$: Let $C$ be a constant, $b$ and $a$ be the values of the predictor $X_i$ (one of the `Lag1` to volume) and

$$\begin{aligned}
\boxed{b > a} &\Rightarrow \beta_i b < \beta_i a \\
&\Rightarrow -\beta_i b > -\beta_i a \\
&\Rightarrow -\beta_i b + C > -\beta_i a + C \\
&\Rightarrow \exp(-\beta_i b + C) > \exp(-\beta_i a + C) \\
&\Rightarrow 1 + \exp(-\beta_i b + C) > 1 + \exp(-\beta_i a + C) \\
&\Rightarrow \frac{1}{1+\exp(-\beta_i b + C)} < \frac{1}{1+\exp(-\beta_i a + C)} \\
&\Rightarrow \boxed{P(Y = 1|X_i = b) < P(Y = 1|X_i = a)}
\end{aligned}$$

where $P(Y = 1|X_i = x) = \frac{1}{1+\exp(-(\beta_i x + \text{other fixed values})}$.

□

(b) With significance level of 5%, write a reduced model for predicting the returns.

*Solution.* Only Lag2 is significant ($p$-value = 0.0296 smaller than $\alpha = 0.05$). The model is

$$\mathbb{P}(Y = 1|X) = \frac{e^{0.2669+0.0584(Lag2)}}{1 + e^{0.2669+0.0584(Lag2)}}.$$

□

3. Suppose we collect data for a group of students in a class with variables $X_1$ = hours studied, $X_2$ = previous GPA, $Y$ = receive an A (1 for yes). We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$ and $\hat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studied for 40 hours with previous GPA of 3.5 gets an A in the class. [Answer: 0.3775]

*Solution.* For $X = (40, 3.5)$,

$$\mathbb{P}(Y = 1|X) = \frac{e^{-6+0.05X_1+X_2}}{1 + e^{-6+0.05X_1+X_2}} = \frac{1}{1 + e^{-(-6+0.05(40)+3.5)}} = 0.3775.$$

□

(b) How many hours would the student in (a) need to study to have 50% chance of getting an A in the class? [Answer: 50]

*Solution.*

$$0.5 = \frac{e^{-6+0.05X_1+3.5}}{1 + e^{-6+0.05X_1+3.5}} \Rightarrow X_1 = 50 \text{ hours}$$

□

4. Suppose that the **Default** dataset is depending on four predictors, `Balance`, `Income`, `Student` and `City`. The results from logistic regression is shown below.

| | Coefficient | Std. error | $Z$-statistic | $P$-value |
|---|---|---|---|---|
| Intercept | -10.8690 | 0.4923 | -22.08 | $< 0.0001$ |
| Balance | 0.0057 | 0.0002 | 24.74 | $< 0.0001$ |
| Income | 0.0030 | 0.0082 | 0.37 | 0.7115 |
| Student [Yes] | -0.6468 | 0.2362 | -2.74 | 0.0062 |
| City_B | 0.1274 | 0.0136 | 10.52 | 0.0003 |
| City_C | 0.0331 | 0.0087 | 5.64 | 0.0011 |

(a) Compare the odds and probability of default between a customer with balance 10,000 and 5,000.

*Solution.*

$$\frac{e^{0.0057(10000)}}{e^{0.0057(5000)}} = 2.3845 \times 10^{12}.$$

The odds of default for a customer with balance 10,000 is $2.3845 \times 10^{12}$ times of the odds of default for a customer with balance 5,000. Hence, the probability of default for the customer with balance 10,000 will be higher. □

(b) Compare the odds and probability of default between a student and a non-student.

*Solution.*

$$e^{-0.6468} = 0.5237$$

The odds of default for a student is 0.5237 times of the odds of default for a non-student. Hence, the probability of default for a student will be lower. □

(c) Compare the odds and probability of default among different cities. [Hint: To "compare" two odds, the best way is to find the odds ratio.]

*Solution.* For City B vs City A:

$$e^{0.1274} = 1.1359$$

The odds of default for City B is 1.1359 times of the odds of default for City A. Hence, the probability of default for City B will be higher.
For City C vs City A:

$$e^{0.0331} = 1.0337$$

The odds of default for City C is 1.0337 times of the odds of default for City A. Hence, the probability of default for City C will be higher.
For City B vs City C:

$$\frac{e^{0.1274}}{e^{0.0331}} = 1.0989$$

The odds of default for City B is 1.0989 times of the odds of default for City C. Hence, the probability of default for City B will be higher.
Comparing all cities, the probability of underprice:

$$\text{City A} < \text{City C} < \text{City B}$$

□

5. (Final Exam Jan 2023, Q2) In a study of the Australian New South Wales Electricity Market, prices are not fixed and are affected by demand and supply of the market. They are set every five minutes. Suppose the collected data $D$ has three attributes `day`, `period`, `transfer` and one output `class` described below:

- $X_1$ =`day`: day of the week, 1–7;
- $X_2$ =`period`: time of the measurement, 1–48, in half hour intervals over 24 hours. It is normalised to between 0 and 1;
- $X_3$ =`transfer`: the scheduled electricity transfer between two Australian states, normalised between 0 and 1; and

- $Y$ =`class`: the change of the price (UP=1 or DOWN=0).

(a) When a logistic regression model is trained with the data, the analysis result below is generated.

```
Call:
glm(formula = class ~ day + period + transfer,
    family = binomial, data = d.f[idx, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5100  -1.0588  -0.8236   1.2279   1.8320

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.207724   0.053076   3.914 9.09e-05 ***
day         -0.054251   0.005844  -9.284  < 2e-16 ***
period       0.956151   0.039873  23.980  < 2e-16 ***
transfer    -1.569139   0.077992 -20.119  < 2e-16 ***
---
Signif:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 43258  on 31716  degrees of freedom
Residual deviance: 42079  on 31713  degrees of freedom
AIC: 42087
```

i. Write down the mathematical expression of the logistic regression model for the three attributes and the output `class`. (4 marks)

*Solution.* The mathematical expression of the logistic regression model is

$$P(Y=1|X_1,X_2,X_3)=\frac{1}{1+\exp(-(0.20772-0.05425X_1+0.95615X_2-1.56914X_3))}.$$

[4 marks]

Average: 3.32 / 4 marks in Jan 2023; 10% below 2 marks. □

ii. Calculate the conditional probability of UP when the day is 2, the period is 0.042553 and the transfer is 0.414912 based on the logistic regression model. (6 marks)

*Solution.* First, we calculate

$$\begin{aligned}\beta^T\boldsymbol{x} &= 0.207724-0.054251\times 2+0.956151\times 0.042553\\ &\quad -1.569139\times 0.414912=-0.5111455\end{aligned}$$

[4 marks]

The probability of UP is

$$\begin{aligned}&P(Y=1|X_1=2,X_2=0.042553,X_3=0.414912)\\ &=\frac{1}{1+\exp(0.5111455)}=0.374925\end{aligned}$$

[2 marks]

Average: 5.60 / 6 marks in Jan 2023; 3% below 3 marks. □

iii. Calculate the odds ratio for the electricity price going UP to the price going DOWN given the day 2 is changed to day 4. (4 marks)

*Solution.*

$$\frac{odds(day=4)}{odds(day=2)}=\exp(-0.054251*(4-2))=\exp(-0.108502)=0.8971771$$

[4 marks]

Average: 2.32 / 4 marks in Jan 2023; 38% below 2 marks. □

6. (Final Exam Jan 2024 Sem, Q2) When a bank receives a loan application, the bank has to make a decision whether to go ahead with the loan approval or not based on the applicant's profile. Two types of risks are associated with the bank's decision:

   - If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank;
   - If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank.

   To minimise loss from the bank's perspective, the bank needs a predictive model regarding who to give approval of the loan and who not to based on an applicant's demographic and socio-economic profiles.

   Suppose the response variable `Y` is 0 when the loan is approved and is 1 when the loan is not approved. Suppose the features of the data are listed below:

   - $X_1$ (categorical): Status of existing checking account (A11, A12, A13, A14);
   - $X_2$ (integer): Duration in months
   - $X_3$ (integer): Credit amount
   - $X_4$ (integer): Instalment rate in percentage of disposable income
   - $X_5$ (binary): foreign worker (yes, no)

   (a) When the data is trained with a logistic regression model, the statistical estimates below are obtained:

```
Call:
glm(formula = Y ~ ., family = binomial, data = d.f.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8613  -0.7239  -0.5115   0.9647   2.0657

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.561e-01  9.771e-01   0.364 0.715529
X1A12       -6.719e-01  6.889e-01  -0.975 0.329365
X1A13       -1.792e+01  2.795e+03  -0.006 0.994884
X1A14       -2.275e+00  6.754e-01  -3.369 0.000755 ***
X2           3.834e-02  3.039e-02   1.262 0.207052
X3          -1.965e-05  1.406e-04  -0.140 0.888871
X4          -1.336e-01  2.634e-01  -0.507 0.612044
X5no        -1.733e+01  1.935e+03  -0.009 0.992854
---
Signif.:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   i. Write down the mathematical expression of the logistic regression model in the conditional probability form. (4 marks)

*Solution.* The mathematical expression of the logistic regression model is

$$P(Y=1|\mathbf{x}) = \frac{1}{1+\exp(-\beta\cdot\mathbf{x})} \qquad \text{[2 marks]}$$

where

$$\beta\cdot\mathbf{x} = 0.3561 - 0.6719x_1^{A12} - 17.92x_1^{A13} - 2.275x_1^{A14} + 0.03834x_2 - 1.965\times10^{-5}x_3 - 0.1336x_4 - 17.33x_5 \qquad \text{[2 marks]}$$

Average: 3.43 / 4 marks in Jan 2024; 3.64% below 2 marks. □

ii. Calculate the conditional probability of $Y = 1$ and the conditional probability of $Y = 0$ for a foreign worker when the status of existing checking account of the customer is A11, the duration is 6 months, the credit amount is 1169 and the instalment rate of disposable income is 4%. (6 marks)

*Solution.* We tabulate the information for calculation:

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $\beta \cdot \mathbf{x}$ |
|---|---|---|---|---|---|---|
| | A11 | 6 | 1169 | 4 | yes | |
| | 0 | $3.834 \times 10^{-2}$ | $-1.965 \times 10^{-5}$ | $-0.1336$ | 0 | |
| 0.3561 | 0 | 0.23004 | $-0.02297085$ | $-0.5344$ | 0 | 0.02876915 |

. . . . . [5 marks]

Therefore,

$$P(Y = 1|X) = \frac{1}{1 + \exp(-(0.02876915))} = 0.5071918 \qquad \text{[0.5 mark]}$$

$$P(Y = 0|X) = 1 - 0.5071918 = 0.4928082 \qquad \text{[0.5 mark]}$$

Average: 5.03 / 6 marks in Jan 2024; 1.82% below 3 marks. □