

Tut 3: Logistic Regression

Feb 2025

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

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

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

$$\mathbb{P}(Y = 1|\mathbf{X}) = \frac{1}{1 + \exp(-(\beta_0 + \dots + \beta_i^{(2)} x_{i.\text{level}2} + \dots + \beta_i^{(K)} x_{i.\text{level}K} + \dots))}$$

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

$$\begin{aligned} \text{Odds} &= \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)} = \frac{\mathbb{P}(Y = 1)}{1 - \mathbb{P}(Y = 1)} = \frac{\frac{\exp(\dots)}{\exp(\dots)+1}}{1 - \frac{\exp(\dots)}{\exp(\dots)+1}} \\ &= \frac{\exp(\dots)}{\exp(\dots) + 1 - \exp(\dots)} = \exp(\dots) = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p). \end{aligned}$$

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

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

Odds Ratio for “one-hot-encoded” categorical value:

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

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, β_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

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

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

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

□

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 2.3845×10^{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. \square

- (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. \square

- (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. Using City A as a reference, we have

$$\frac{\text{odds}(\text{City B})}{\text{odds}(\text{City A})} = e^{0.1274} = 1.1359$$

$$\frac{\text{odds}(\text{City C})}{\text{odds}(\text{City A})} = e^{0.0331} = 1.0337$$

This implies

$$\text{odds}(\text{City A}) < \text{odds}(\text{City C}) = 1.0337\text{odds}(\text{City A}) < \text{odds}(\text{City B}) = 1.1359\text{odds}(\text{City A})$$

which implies

$$P(\text{City A}) < P(\text{City C}) < P(\text{City B})$$

\square

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:

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

$$\beta^T \mathbf{x} = 0.207724 - 0.054251 \times 2 + 0.956151 \times 0.042553 - 1.569139 \times 0.414912 = -0.5111455$$

[4 marks]

The probability of UP is

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

[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.

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

[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

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

where

$$\beta \cdot \mathbf{x} = 0.3561 - 0.6719x_1^{A12} - 17.92x_1^{A13} - 2.275x_1^{A14} + 0.03834x_2 - 1.965 \times 10^{-5}x_3 - 0.1336x_4 - 17.33x_5 \quad [2 \text{ marks}]$$

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	3.834×10^{-2}	-1.965×10^{-5}	-0.1336	0	
0.3561	0	0.23004	-0.02297085	-0.5344	0	0.02876915

..... [5 marks]
 Therefore,

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

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

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

7. (Final Exam May 2024 Sem, Q2) Consider a heart attack classification dataset with a binary output with 0 and 1 representing less chance and more chance of getting heart attack respectively. The features of the dataset are

- **sex** : sex of the patient (0 for female, 1 for male);
- **cp** : chest pain type with the value 0 representing “typical angina”, the value 1 representing “atypical angina”, the value 2 representing “non-anginal pain”, the value 3 representing “asymptomatic”;
- **trtbps** : resting blood pressure (in mm Hg);
- **chol** : cholesterol in mg/dl fetched via BMI sensor;
- **thalachh** : maximum heart rate achieved

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

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.381684   1.695896  -0.225  0.821930
sex1         -2.010528   0.381541  -5.269  1.37e-07 ***
cp1          2.134406   0.476738   4.477  7.57e-06 ***
cp2          2.095930   0.366266   5.722  1.05e-08 ***
cp3          2.142175   0.585113   3.661  0.000251 ***
trtbps       -0.026184   0.008943  -2.928  0.003412 **
chol         -0.007099   0.003082  -2.303  0.021252 *
thalachh     0.040776   0.008257   4.939  7.86e-07 ***
---
Signif codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 266.49  on 295  degrees of freedom
AIC: 282.49

Number of Fisher Scoring iterations: 5
```

(i) Write down the mathematical expression of the logistic regression model in the proper conditional probability form and then use it to predict the probability of more chance of getting heart attack for a female who has non-anginal chest pain, the resting blood pressure is 130 mm Hg, the cholesterol is 275 mg/dl and the maximum heart rate achieved is 139. (10 marks)

Solution. The mathematical expression of the logistic regression model is

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

where

$$\begin{aligned} \beta \cdot \mathbf{x} = & -0.381684 - 2.010528\text{sex1} + 2.134406\text{cp1} + 2.095930\text{cp2} + 2.142175\text{cp3} \\ & - 0.026184\text{trtbps} - 0.007099\text{chol} + 0.040776\text{thalachh} \end{aligned} \quad [2 \text{ marks}]$$

We tabulate the information for calculation:

	sex	cp	trtbps	chol	thalachh	$\beta \cdot \mathbf{x}$
$x^* =$	female=0	cp2=1	130	275	139	
-0.381684	-2.010528	2.095930	-0.026184	-0.007099	0.040776	
-0.381684	0	2.095930	-3.40392	-1.952225	5.667864	2.025965

..... [5 marks]

Therefore,

$$P(Y = 1|X = x^*) = \frac{1}{1 + \exp(-(2.025965))} = 0.8835 \quad [1 \text{ mark}]$$

□

- (ii) Calculate the odds ratio of the chest pain types and then compare the probabilities of more chance of getting heart attack for the chest pain types. Justify your answers.

(4 marks)

Solution.

$$\begin{aligned} \frac{\text{odds}(cp = 1)}{\text{odds}(cp = 0)} &= \exp(2.134406) = 8.452025 \\ \frac{\text{odds}(cp = 2)}{\text{odds}(cp = 0)} &= \exp(2.095930) = 8.133001 \quad [3 \text{ marks}] \\ \frac{\text{odds}(cp = 3)}{\text{odds}(cp = 0)} &= \exp(2.142175) = 8.51794 \end{aligned}$$

Since the odds and conditional probabilities preserves the ordering, we have the probabilities of more chance of getting heart attack for the chest pain types satisfying the comparison below:

$$\text{cp} = 0 < \text{cp} = 2 < \text{cp} = 1 < \text{cp} = 3 \quad [0.5 \text{ mark}]$$

or

$$\text{typical angina} < \text{non-anginal pain} < \text{atypical angina} < \text{asymptomatic.} \quad [0.5 \text{ mark}]$$

□