Tut 3: Logistic Regression

Jan 2023

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

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

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. $\dfrac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)} = 0.37 \Rightarrow \mathbb{P}(Y = 1|X) = \dfrac{0.37}{1 + 0.37} = 0.270073$

   $\therefore$ fraction/probability $= 27\%$ $\square$

   (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 $= \dfrac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)} = \dfrac{0.16}{1 - 0.16} = 0.190048$

   $\therefore$ odds $= 19\%$. $\square$

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

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

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 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 a customer with balance 10,000 is 2.3845×10^{12} times of the odds of 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 a student is 0.5237 times of the odds of 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 City B is 1.1359 times of the odds of 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 City C is 1.0337 times of the odds of 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 City B is 1.0989 times of the odds of City C. Hence, the probability of default for City B will be higher.

Comparing all cities, the probability of underprice:

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

□

4. 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 $X = (40, 3.5)$,

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

$\square$

(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.

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

$\square$