Tut 3: Logistic Regression

Jan 2022

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,

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

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%]

   (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%]

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.

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

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.

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

(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.]

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]

   (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]