Predictive Modelling Tutorial 3: Logistic Regression

Dr Liew How Hui

Jan 2021

Tut 3: Logistic Regression

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

Tut 3: Logistic Regression

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

Tut 3: Logistic Regression

Let $k = 2, \ldots, K$. Odds Ratio,

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

Tutorial 4, Q1

(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%

Tutorial 4, Q2

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

Tutorial 4, Q2 (cont)

(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.

Tutorial 4, Q3

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

Tutorial 4, Q3 (cont)

(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.

Note: To “compare” two odds, the best way is to find the odds ratio.

Tutorial 4, Q4

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

FA May 2020 Q2 (a)

The testing dataset of an insurance claim is given in Table 2.1. The variables “gender”, “bmi”, “age_bracket” and “previous_claim” are the predictors and the “claim” is the response.

Table 2.1: The testing data of an insurance claim (randomly sampled with repeated entry).

gender	bmi	age_bracket	previous_claim	claim
female	under_weight	18-30	0	no_claim
female	under_weight	18-30	0	no_claim
male	over_weight	31-50	0	no_claim
female	under_weight	50+	1	no_claim
male	normal_weight	18-30	0	no_claim
female	under_weight	18-30	1	no_claim
male	over_weight	18-30	1	no_claim
male	over_weight	50+	1	claim
female	normal_weight	18-30	0	no_claim
female	obese	50+	0	claim

FA May 2020 Q2 (a) cont

The “gender” is binary categorical data, the “bmi” is a four-value categorical data with values under_weight, normal_weight, over_weight and obese, the “age_bracket” is a three-value categorical data with value “18-30”, “31-50” and “50+”, the “previous_claim” is a binary categorical data with 0 indicating “no previous claim” and 1 indicating “having a previous claim”. The “claim” is a binary response with values “no_claim” (negative class, with value 1) and “claim” (positive class, with value 0).

FA May 2020 Q2 (a) cont

Suppose a logistic regression model is trained and the coefficients are stated in Figure 2.2.

Figure 2.2: The coefficients of the logistic regression based on an insurance claim data.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.1361     0.2990  10.489  < 2e-16 ***
gendermale         -0.3343     0.1753  -1.908  0.05644 .
bmiobese           -1.9495     0.2821  -6.910 4.86e-12 ***
bmiover_weight     -1.0563     0.2629  -4.017 5.89e-05 ***
bmiunder_weight    -0.8424     0.2606  -3.232  0.00123 **
age_bracket31-50   -0.2875     0.2313  -1.243  0.21382
age_bracket50+     -1.2133     0.2241  -5.414 6.18e-08 ***
previous_claim1    -0.9505     0.1763  -5.392 6.96e-08 ***
---
Signif. :  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FA May 2020 Q2 (a) cont

Write down the **mathematical formula** of the logistic regression model and then use it to **predict** the "claim" of the insurance data in Table 2.1 as well as **evaluating** the performance of the model by calculating the confusion matrix, accuracy, sensitivity, specificity, PPV, NPV of the logistic model. [**Note**: The default cut-off is 0.5] (4 marks)