Tut 4: Logistic Regression (cont)

June 2024

1. (Jan 2022 Final Q2(a)) Given the following results from the analysis of credit card applications approval dataset using logistic regression model.

```
glm(formula=Approved~., family=binomial, data=d.f.train)

Deviance Residuals:
    Min        1Q    Median       3Q       Max
-2.6796   -0.5477   0.2681   0.3316   2.4501

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.1379649  0.5744168   5.463 4.68e-08 ***
Maleb         -0.1758676  0.3229541  -0.545   0.5861
Age            0.0001318  0.0142338   0.009   0.9926
Debt           0.0042129  0.0298740   0.141   0.8879
YearsEmployed -0.1023132  0.0582368  -1.757   0.0789 .
PriorDefaultt -3.6614227  0.3659226 -10.006  < 2e-16 ***
Employedt     -0.2500687  0.4013495  -0.623   0.5332
CreditScore   -0.1098142  0.0644360  -1.704   0.0883 .
ZipCode        0.0011958  0.0009540   1.253   0.2100
Income        -0.0004544  0.0001966  -2.311   0.0209 *
---
Signif.:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 625.90  on 454  degrees of freedom
Residual deviance: 294.33  on 445  degrees of freedom
  (27 observations deleted due to missingness)
AIC: 314.33
```

where the output `Approved` is either positive (represented as 0) and negative (represented as 1) and the features

- `Male` is categorical with `a`=Female, `b`=Male;
- `PriorDefault` is categorical with `f`=false, `t`=true;
- `Employed` is categorical with `f`=false, `t`=true;
- `Age`, `Debt`, `YearsEmployed`, `CreditScore`, `ZipCode`, `Income` are continuous variables.

(a) Write down the mathematical expression of the logistic model for the given data with the coefficient values rounded to 4 decimal places. (4 marks)

Solution. The logistic model is

$$\mathbb{P}(\texttt{Approved} = 1|\boldsymbol{X}) = \frac{1}{1+e^{-(3.1380+\boldsymbol{w}^T\boldsymbol{X})}} \qquad \text{[1.5 mark]}$$

$$\begin{aligned}\boldsymbol{w}^T\boldsymbol{X} = &-0.1759\,\texttt{Male} + 0.0001\,\texttt{Age} + 0.0042\,\texttt{Debt} - 0.1023\,\texttt{YearsEmployed}\\ &- 3.6614\,\texttt{PriorDefault} - 0.2501\,\texttt{Employed} - 0.1098\,\texttt{CreditScore}\\ &+ 0.0012\,\texttt{ZipCode} - 0.0005\,\texttt{Income}\end{aligned}$$

[2.5 marks]

□

(b) By calculating the probability of the credit card application being approved for a male of age 22.08 with a debt of 0.83 unit who has been employed for 2.165 years with no prior default and is currently unemployed, has a credit score 0 and a zip code 128 with income 0, find the **probability** of credit card applications approval and determine if the approval is positive or negative (using the cut-off of 0.5). (7 marks)

Solution. First, we calculate

$$\begin{aligned}\boldsymbol{w}^T\boldsymbol{X} &= -0.1759\,(1) + 0.0001\,(22.08) + 0.0042\,(0.83) - 0.1023\,(2.165)\\ &\quad - 3.6614\,(0) - 0.2501\,(0) - 0.1098\,(0)\\ &\quad + 0.0012\,(128) - 0.0005\,(0)\\ &= -0.2380855\end{aligned}$$

[4 marks]

The probability of the credit card application being 'negatively' approved,

$$\mathbb{P}(\texttt{Approved} = 1|\boldsymbol{X}) = \frac{1}{1 + \exp(-\underbrace{(3.1380 - 0.2380855)}_{2.899914})} = 0.9478$$

[2 marks]

Since the probability is more than 0.5, the approval is **negative**. … [1 mark] □

(c) Calculate the odds ratio for the approval being negative with the prior default to be true against the prior default to be false. Infer the likelihood of getting a negative approval based on the prior default. (6 marks)

Solution. The odds ratio for the approval with respect to prior default is

$$\frac{\frac{\mathbb{P}(\texttt{Approved}=1|\texttt{PriorDefault}=t)}{1-\mathbb{P}(\texttt{Approved}=1|\texttt{PriorDefault}=t)}}{\frac{\mathbb{P}(\texttt{Approved}=1|\texttt{PriorDefault}=f)}{1-\mathbb{P}(\texttt{Approved}=1|\texttt{PriorDefault}=f)}} = \frac{\exp(-3.6614227 \times 1)}{\exp(-3.6614227 \times 0)} = 0.02569593$$

[4 marks]

Someone with a prior default has a lower likelihood to get a negative approval compare to someone without a prior default. .... [2 marks] □

2. (May 2020 Final 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

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).

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
```

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)

Solution. Let X be all the dummy variables associated with the four predictors and Y be the response variable Y. The mathematical formula is

$$\mathbb{P}(Y=1|X) = \frac{1}{1+\exp(-(3.1361+\beta^T X))} \qquad [0.4 \text{ mark}]$$

where

$$\begin{aligned}\beta^T X = &-0.3343 \cdot \text{male} - 1.9495 \cdot \text{obese} - 1.0563 \cdot \text{overweight} - 0.8424 \cdot \text{underweight} \\ &- 0.2875 \cdot \text{age31} - 1.2133 \cdot \text{age50} - 0.9505 \cdot \text{prv.claim.1}\end{aligned}$$

[0.6 mark]

The prediction of the testing data is given below:

male	obese	over.wt	under.wt	age31	age50	prv.claim.1	prob	$\widehat{Y}$	Y
0	0	0	1	0	0	0	0.9083545	1	no_claim
0	0	0	1	0	0	0	0.9083545	1	no_claim
1	0	1	0	1	0	0	0.8112102	1	no_claim
0	0	0	1	0	1	1	0.5324324	1	no_claim
1	0	0	0	0	0	0	0.9427711	1	no_claim
0	0	0	1	0	0	1	0.7930248	1	no_claim
1	0	1	0	0	0	1	0.6889022	1	no_claim
1	0	1	0	0	1	1	0.3969113	0	claim
0	0	0	0	0	0	0	0.9583570	1	no_claim
0	1	0	0	0	1	0	0.4933065	0	claim

. . . . [2 marks]

The confusion matrix is as follows

	claim (0)	no_claim (1)
predict 0	2	0
predict 1	0	8

.... [0.5 mark]

The performance metrics are
Accuracy : 1

Sensitivity : 1

Specificity : 1

Pos Pred Value : 1

Neg Pred Value : 1 .... [0.5 mark]

□

3. (Final Exam May 2023, Q2) A bank customer churn dataset contains information on the customers:

- `Creditscore`: the score represent the summary of a bank customer credit history and indicate the likelihood of repaying borrowed funds;
- `Geography`: a categorical feature with values France, Germany, Spain;
- `Gender`: a binary categorical feature with values Female, Male;
- `Age`: the age of the customer (integer value);
- `Balance`: the amount a customer have in their account;
- `NumOfProducts`: the number of products a bank customer purchased through the bank;
- `IsActiveMember`: a categorical variable indicating whether a customer is active (1) or inactive (0);

The response variable `Exited` shows if a customer has been churned ($Y = 1$) or not ($Y = 0$).

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

```
Call:
glm(formula = Exited ~ ., family = binomial, data = D.train)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.335e+00  3.274e-01 -10.188  < 2e-16 ***
CreditScore      -7.811e-04  3.931e-04  -1.987   0.0469 *
GeographyGermany  7.888e-01  9.542e-02   8.266  < 2e-16 ***
GeographySpain   -2.094e-02  1.002e-01  -0.209   0.8344
GenderMale       -5.206e-01  7.700e-02  -6.761 1.37e-11 ***
Age               7.211e-02  3.683e-03  19.581  < 2e-16 ***
Balance           3.061e-06  7.318e-07   4.183 2.88e-05 ***
NumOfProducts    -1.413e-01  6.723e-02  -2.101   0.0356 *
IsActiveMember1  -1.062e+00  8.151e-02 -13.024  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5053.1  on 4998  degrees of freedom
Residual deviance: 4285.5  on 4990  degrees of freedom
AIC: 4303.5
```

i. Write down the mathematical expression of the logistic regression model for all the features and the response `Exited` denoted by Y. (4 marks)

Solution. Let X_1 denote `Creditscore`, X_2^G denote the dummy variable `GeographyGermany`, X_2^S denote the dummy variable `GeographySpain`, X_3 denote the dummy variable `GenderMale`, X_4 denote `Age`, X_5 denote `Balance`, X_6 denote `NumOfProducts` and X_7 denote the dummy variable `IsActiveMember1`.
The mathematical expression of the logistic regression model is

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

where

$$\begin{aligned}\beta \cdot \mathbf{x} = &-3.335 - 7.811 \times 10^{-4} X_1 + 0.7888 X_2^G - 0.02094 X_2^S - 0.5206 X_3 \\ &+ 0.07211 X_4 + 3.061 \times 10^{-6} X_5 - 0.1413 X_6 - 1.062 X_7\end{aligned}$$

[2 marks]

□

ii. Calculate the conditional probability of churned for a male customer of age 36 and geographically located in Spain with a `CreditScore` 749, a zero `Balance`, having two products and is not an active member. (6 marks)

Solution. We tabulate the information for calculation:

	CreditScore	Spain	Male	Age	Balance	#Products	IsActiveMember
	749	1	1	36	0	2	0
-3.335	-7.811×10^{-4}	-0.02094	-0.5206	0.07211	3.061×10^{-6}	-0.1413	-1.062
-3.335	-0.5850	-0.02094	-0.5206	2.5960	0	-0.2826	0

which sums to -2.14814. .... [5 marks]
Therefore, the conditional probability of churned is

$$P(Y = 1|X) = \frac{1}{1 + \exp(-(-2.14814))} = 0.104505 \qquad \text{[1 mark]}$$

□

iii. Compare the odds and probability of churned among different geographies using the notion of odds ratio and logistic regression model. (4 marks)

Solution. The odds ratio of Germany against France is

$$\frac{odds(Y = 1|X_2 = \text{Germany})}{odds(Y = 1|X_2 = \text{France})} = \exp(7.888 \times 10^{-1}) = 2.200754 > 1 \qquad \text{[1 mark]}$$

The odds ratio of Spain against France is

$$\frac{odds(Y = 1|X_2 = \text{Spain})}{odds(Y = 1|X_2 = \text{France})} = \exp(-2.094 \times 10^{-2}) = 0.979278 < 1 \qquad \text{[1 mark]}$$

These imply the comparison of odds of churned among different geographies:

$$odds(Y = 1|X_2 = \text{Spain}) < odds(Y = 1|X_2 = \text{France}) < odds(Y = 1|X_2 = \text{Germany})$$

[1 mark]

which then implies the comparison of probabilities of churned among different geographies:

$$P(Y = 1|X_2 = \text{Spain}) < P(Y = 1|X_2 = \text{France}) < P(Y = 1|X_2 = \text{Germany})$$

[1 mark]

□