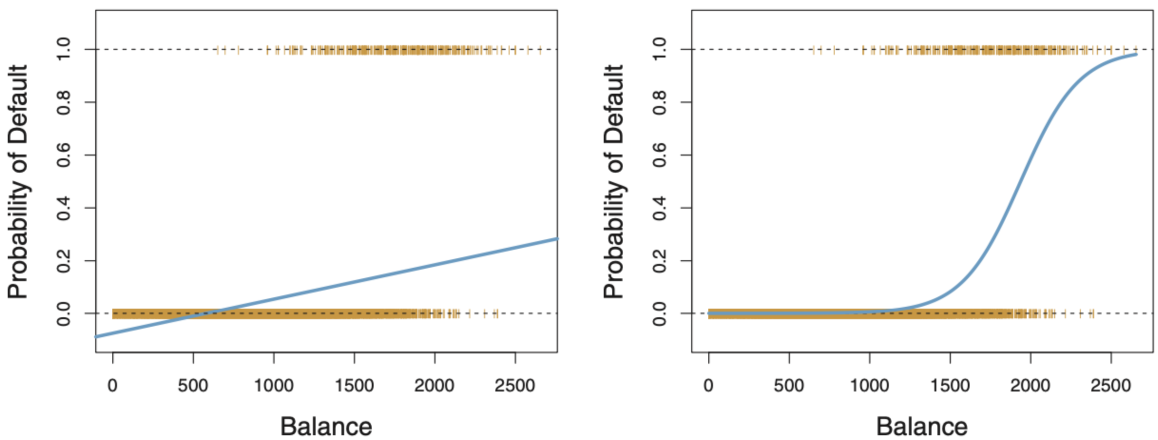
4.3 Logistic Regression

Zongyi Liu

2023-05-19

4.3 Logistic Regression

In the data set, there are two categories, and there are some probability to assign each point to particular category. In the plot, the orange ticks indicate the 0/1 values coded for Yes or No to be in the category.



In the `Default` data, logistic regression models the probability of default. For example, the probability of default given `balance` can be written as

$$Pr(default = Yes|balance)$$

The value of Pr will be ranging from 0 to 1.

4.3.1 The Logistic Model

If we use the linear regression as below, to model the relationship,

$$p(X) = \beta_0 + \beta_1 X.$$

We would get a result similar to the left-hand panel of the figure above: for balances close to zero we predict a negative probability of default; if we were to predict for very large balances, we would get values bigger than 1. To avoid this disadvantage, we would use the logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

which can be manipulated into using maximum likelihood method

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

The quantity $p(X)/(1-p(X))$ is called the odds, and can take on any value between 0 and infinity.

Taking the logarithm of both sides, we can get

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

The left-hand side is called the **log-odds**, or **logit**. In the logistic regression model, there is a logit is linear in X.

In linear model, β_1 is associated with one-unit increase in X, where as in logistic model, one-unit increase in X would cause the odds to increase by e_1^{β}.

4.3.2 Estimating the Regression Coefficients

The coefficients β_0 and β_1 are unknown, and must be estimated using training data. Here we will use the **likelihood function** to estimate:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ will be chosen to maximize this function.

4.3.3 Making Predictions

Once the coefficients have been estimated, it is simple to compute the probability of `default` for any given credit card balance. For example, we can predict the default probability for an individual with a `balance` of $1,000:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576,$$

Which is very small.

We can also use the qualitative predictors with the logistic regression model using dummy variable approach, here student is encoded as 1, and non-student is encoded as 0:

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Here the coefficient associated with the student is positive whereas the p-value is significant, indicating that students tend to have higher default probabilities than non-students.

4.3.4 Multiple Logistic Regression

With the same logic of multiple linear regression, we can generalize the logistic regression as follows:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

p can be written as:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

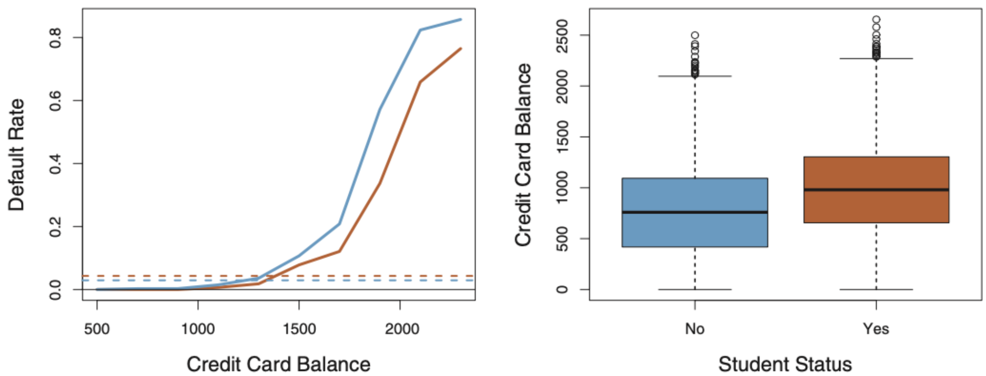
We will still use the maximum likelihood method to estimate $\beta_0, \beta_1, \ldots, \beta_p$.

Example:

	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

Here student status is encoded as a dummy variable `student[Yes]` with 1 for student and 0 for a non-student.

Plots can be used to show the credit card balance for student (orange) and non-student (blue).



By substituting estimates for the regression coefficients from the table above, we can make predictions. For example, a student with a credit card balance of \$1,500 and an income of \$40,000 has an estimated probability of default of:

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058.$$

and a non-student with the same settings has a probability of default:

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105.$$

4.3.5 Logistic Regression for >2 Response Classes

There is a multiple-class extension of the two-class logistic regression model, but it is not to be used all that often.