Classification
Ch4-ISLP; Ch2-IMLP

Zhongjian Lin
University of Georgia

September 8, 2024

Qualitative Output

The linear regression model discussed assumes that the response variable Y is quantitative. But in many situations, the response variable is instead qualitative.

- Eye color.
- Purchase decision.
- Market Entry or not.
- R&D in new product or not.
- Fraudulent transaction.
- ...

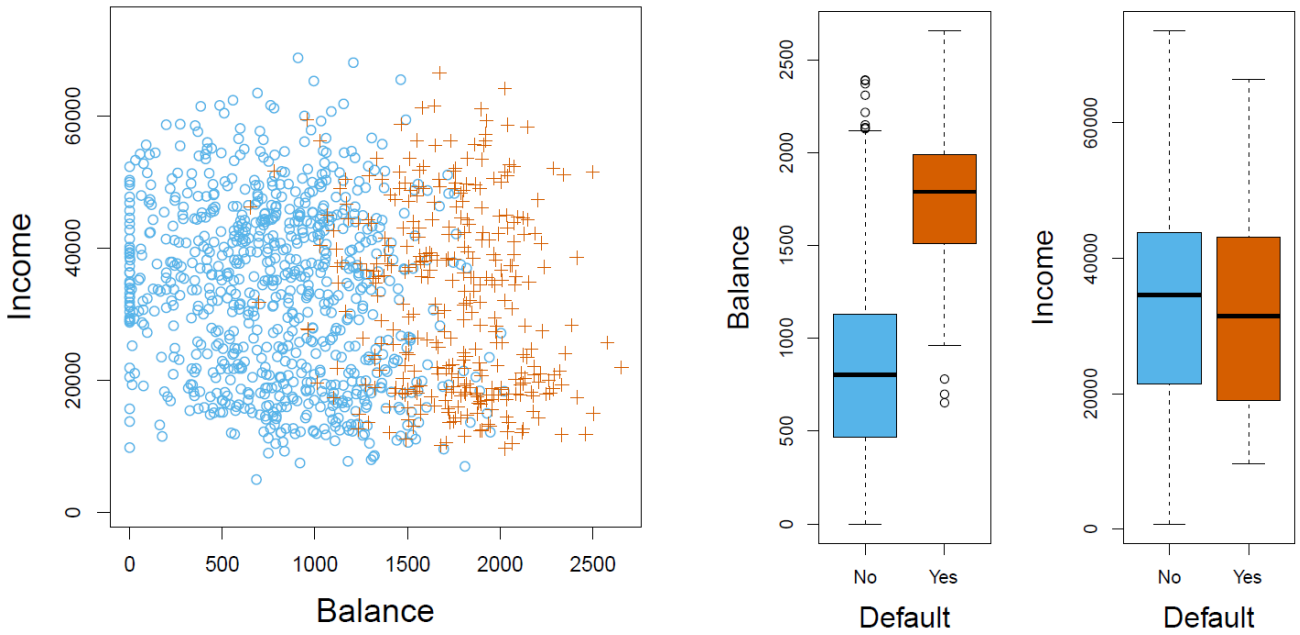
Predicting qualitative responses is a process that is known as classification. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class. On the other hand, often the methods used for classification first predict the probability that the observation belongs to each of the categories of a qualitative variable, as the basis for making the classification.

Classification

In the classification setting we have a set of training observations

$$(x_1, y_1), \cdots, (x_n, y_n)$$

that we can use to build a classifier. We want our classifier to perform well not only on the training data, but also on test observations that were not used to train the classifier.



Why not Linear Regression?

Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms. In this simplified example, there are three possible diagnoses: stroke, drug overdose, and epileptic seizure. We could consider encoding these values as a quantitative response variable, Y, as follows:

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

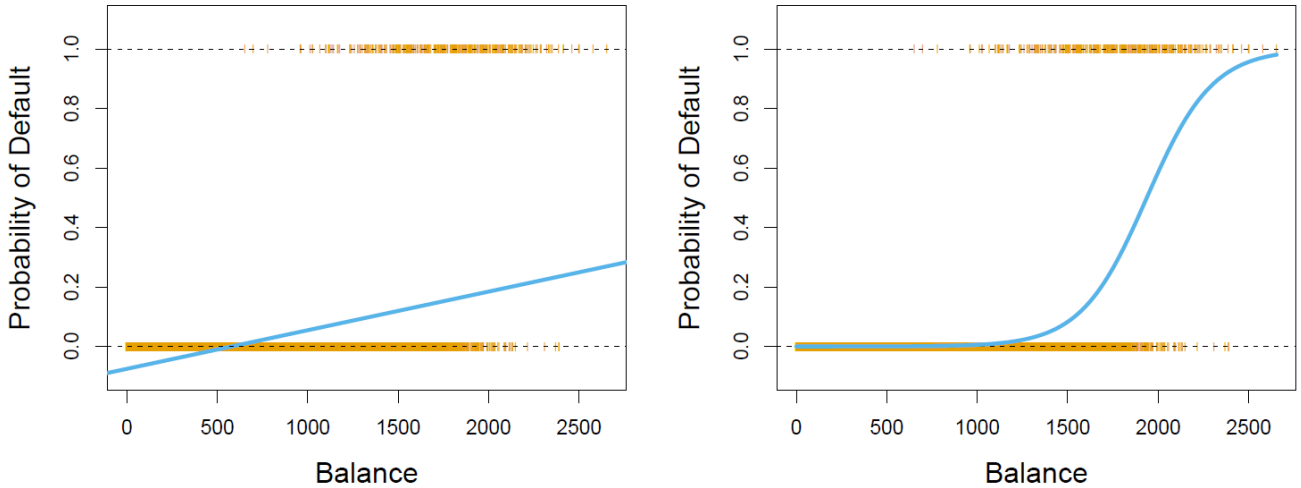
Using this coding, least squares could be used to fit a linear regression model to predict Y on the basis of a set of predictors $X_1, \cdots, X_p$. Though, there is no natural foundation to order in this way.
If the response variable's values did take on a natural ordering, such as mild, moderate, and severe, and we felt the gap between mild and moderate was similar to the gap between moderate and severe, then a 1, 2, 3 coding would be reasonable. Unfortunately, in general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression.

Binary Response

For a binary response with a 0/1 coding, regression by least squares is not completely unreasonable: it can be shown that the $X'\hat{\beta}$ obtained using linear regression is in fact an estimate of $Pr(Y = 1|X)$ in this special case.

$$E(Y|X) = X'\beta + E(\epsilon|X) = X'\beta.$$

However, if we use linear regression, some of our estimates might be outside the [0, 1] interval, making them hard to interpret as probabilities!



- A regression method cannot accommodate a qualitative response with more than two classes;
- A regression method will not provide meaningful estimates of $Pr(Y = 1|X)$, even with just two classes.

Logistic Regression

Instead of having the linear index, $Pr(Y = 1|X) = \beta_0 + \beta_1 X$, we have a small modification to avoid the disparity.

$$p(X) = Pr(Y = 1|X) = F(\beta_0 + \beta_1 X),$$

where $F(\cdot) \in [0, 1]$ is the CDF (cumulative distribution function) of a distribution. When we set $F(\cdot)$ to be the CDF of the standard Logistic distribution, we have the logistic regression

$$Pr(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

Define *odds* as $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$. Values of the odds close to 0 and ∞ indicate very low and very high probabilities, respectively. Odds are traditionally used instead of probabilities in horse-racing, since they relate more naturally to the correct betting strategy. We have *log odds* as

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

In a logistic regression model, increasing X by one unit changes the log odds by β_1.

Estimation

Although we could use (non-linear) least squares to fit the model , the more general method of *maximum likelihood* is preferred, since it has better statistical properties. The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows: we seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual, using Equation (1), corresponds as closely as possible to the individual's observed default status. This intuition can be formalized using a mathematical equation called a likelihood function:

$$l(\beta_0, \beta_1) = \Pi_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i}.$$

The MLE is then

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\max_{\beta \in B} \ l(\beta)$$

or

$$\hat{\beta} = \arg\max_{\beta \in B} \ L(\beta)$$

where $L(\beta) \equiv \log l(\beta)$ is the log likelihood function.

Example: Default

The following table shows the coefficient estimates and related information that result from fitting a logistic regression model on the Default data in order to predict the probability of default=Yes using balance.

	Coefficient	Std. error	z-statistic	p-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Many aspects of the logistic regression output are similar to the linear regression output. For example, we can measure the accuracy of the coefficient estimates by computing their standard errors. The z-statistic plays the same role as the t-statistic in the linear regression output.

Prediction

Once the coefficients have been estimated, we can compute the probability of default $(Y = 1)$ for any given credit card balance $(X = x)$.

$$\hat{p}(x) = \frac{e^{x'\hat{\beta}}}{1 + e^{x'\hat{\beta}}}.$$

For an individual with a balance of \$1, 000 is

$$\hat{p}(1000) = \frac{e^{-10.6513+5.5}}{1 + e^{-10.6513+5.5}} = 0.00576.$$

What is $\hat{p}(2000)$?

Multiple Logistic Regression

In general, we have $X = (X_1, \cdots, X_p)'$ as inputs and

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}. \tag{2}$$

We construct the likelihood function or the log likelihood function similarly and have

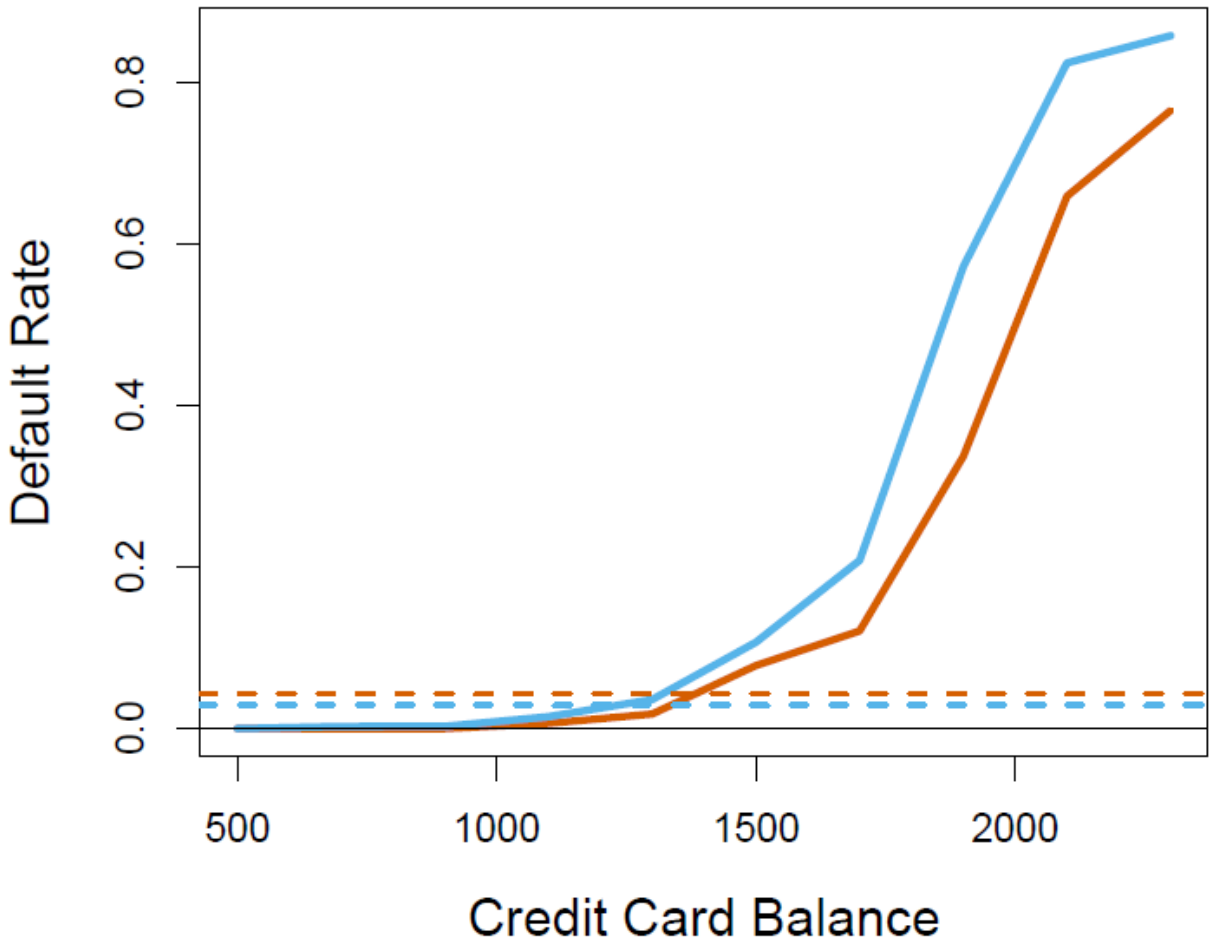
$$\hat{\beta} = \arg\max_{\beta \in B} \; L(\beta) = \sum_{i=1}^{n} y_i \log[p(x_i)] + (1-y_i)\log[1-p(x_i)]$$

Default

$$default = 1\{\beta_0 + \beta_1 balance + \beta_2 income + \beta_3 student + \epsilon \geq 0\}$$

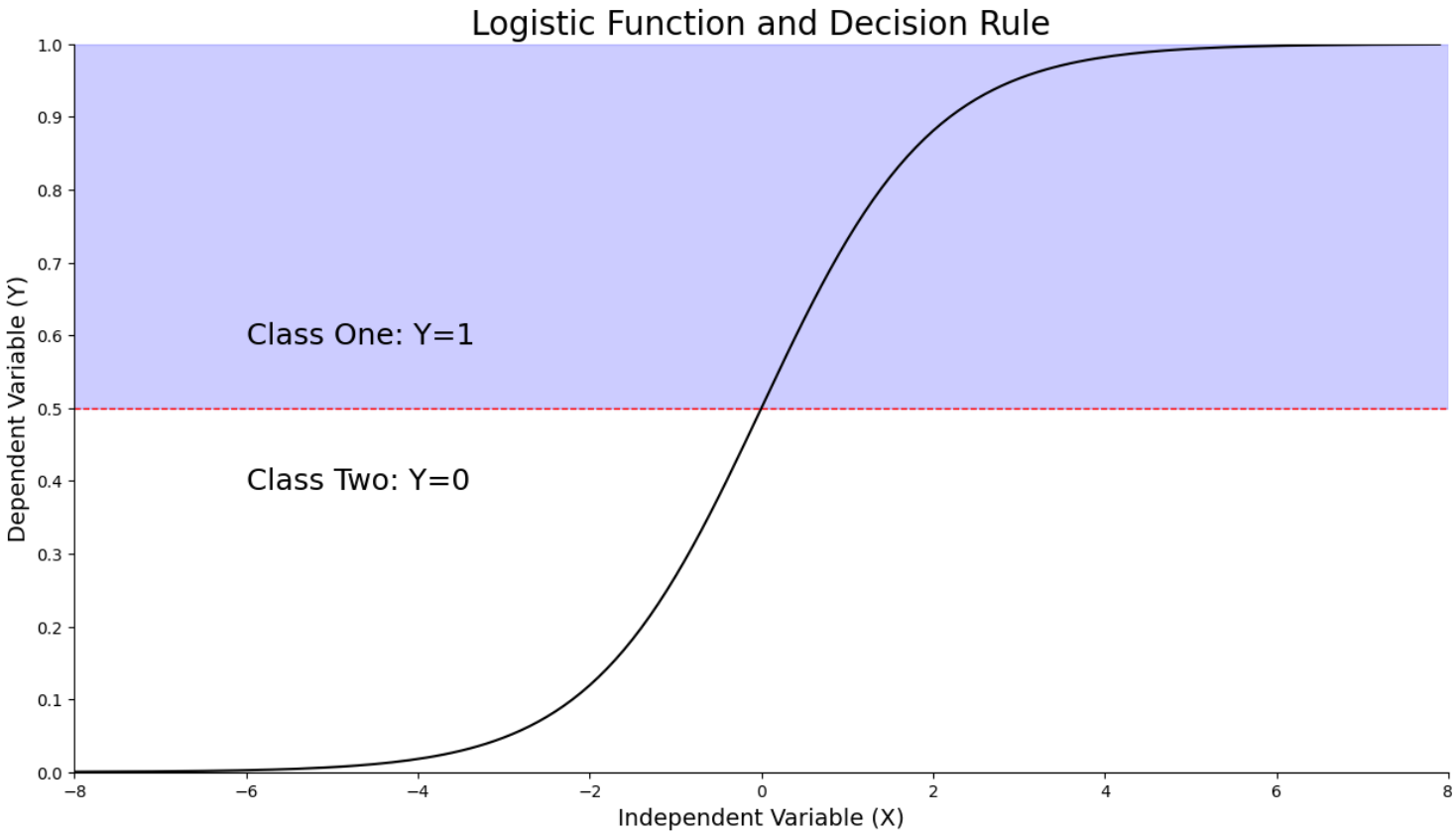
	Coefficient	Std. error	z-statistic	p-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student	-0.6468	0.2362	-2.74	0.0062

Default


Default Rate
0.0
0.2
0.4
0.6
0.8
500
1000
1500
2000
Credit Card Balance

Prediction

The classifier gives us a set of outputs or classes based on probability. Here, we predict the outcome variable is true whenever $P(y = 1|X) \geq 0.5$.



Multinomial Logistic Regression

We sometimes wish to classify a response variable that has more than two classes. For example, we had three categories of medical condition in the emergency room: stroke, drug overdose, epileptic seizure. It turns out that it is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes. This extension is sometimes known as *multinomial logistic regression*. To do this, we first select a single multinomial logistic regression class to serve as the baseline; without loss of generality, we select the Kth class for this role. Then we have the model

$$Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}, k = 1, \cdots, K-1. \quad (3)$$

Furthermore, we have

$$\log\left[\frac{Pr(Y = k|X = x)}{Pr(Y = K|X = x)}\right] = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p.$$

In a setting with K classes, so that we assign an observation to the class that maximizes $Pr(Y = k|X = x)$.