Statistical Method
Logistic Regression

I-Chen Lee, STAT, NCKU

Sections 4.2 and 4.3
Gareth et al. (2021). An Introduction to Statistical Learning with Applications in R.

Nov 28, 2023

Overview

Example: Default dataset

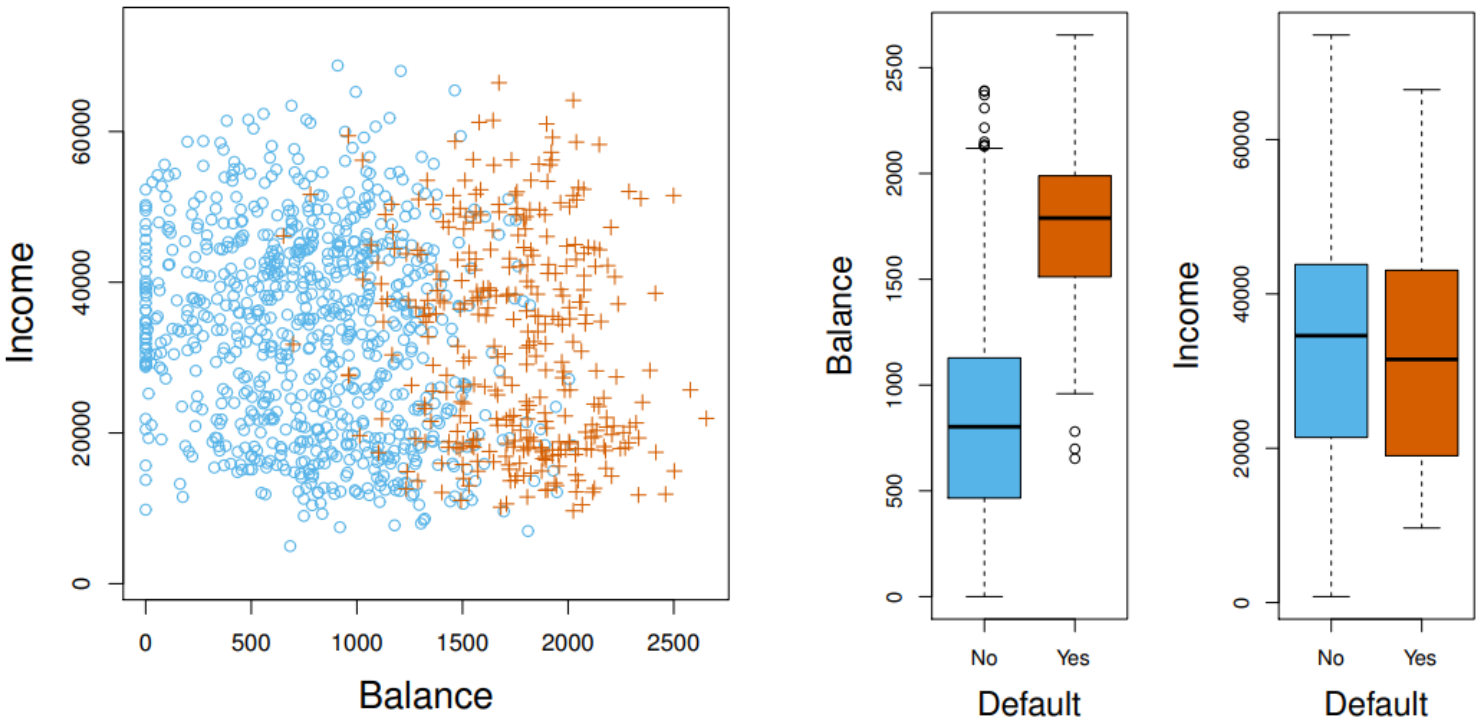


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

Research questions

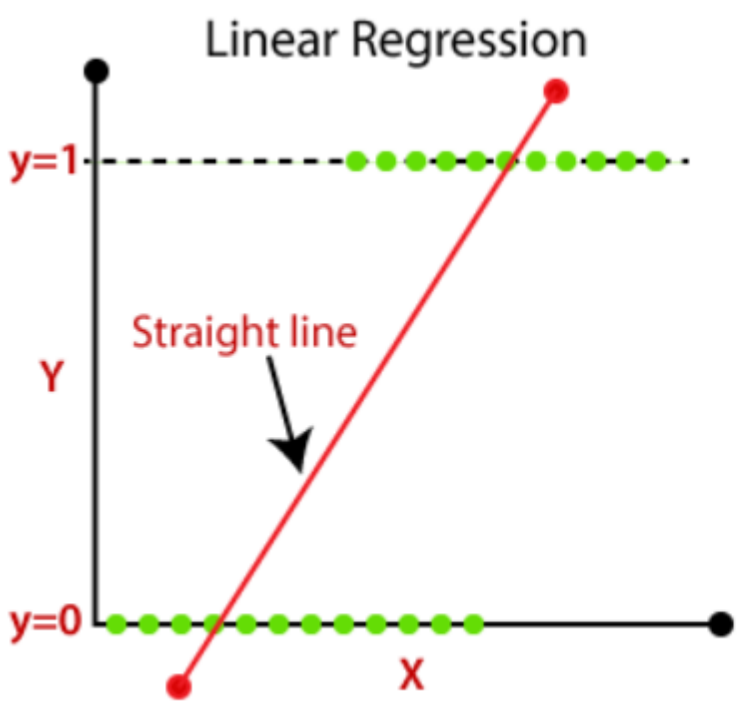
Define the dependent variable (Y) and independent variables (X).

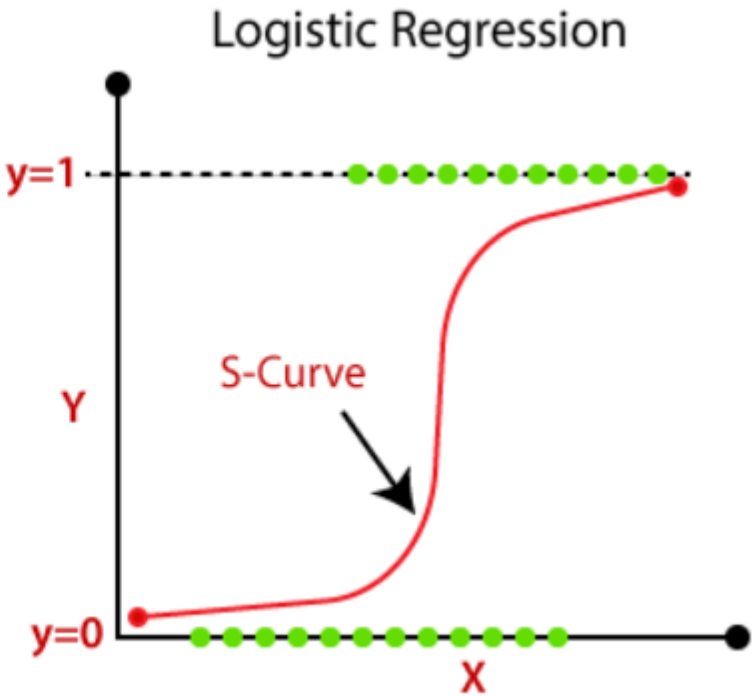
- Relationship between the predictor balance and the response default.

- What causes more on the results of default?

Why not linear regression?

If the response (outcome, dependent variable) is categorical, how to build a model with given covarites (independent variables)?





Example: Default dataset

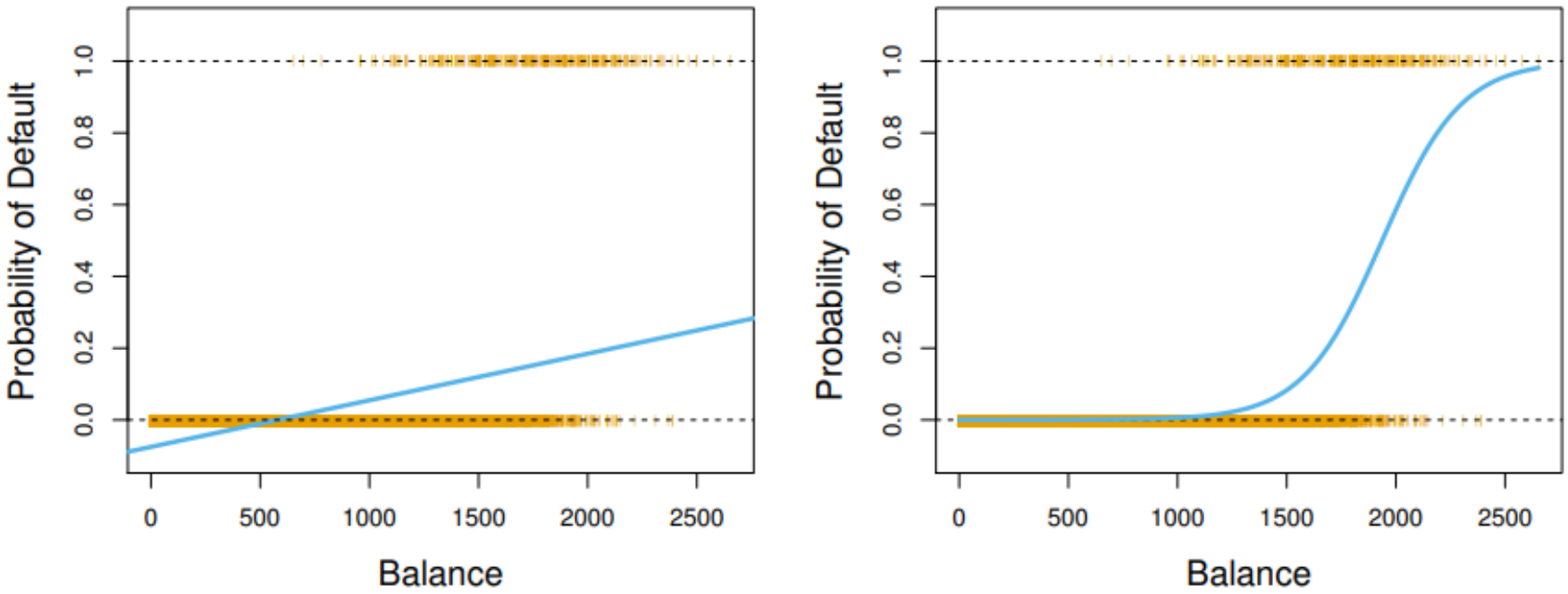
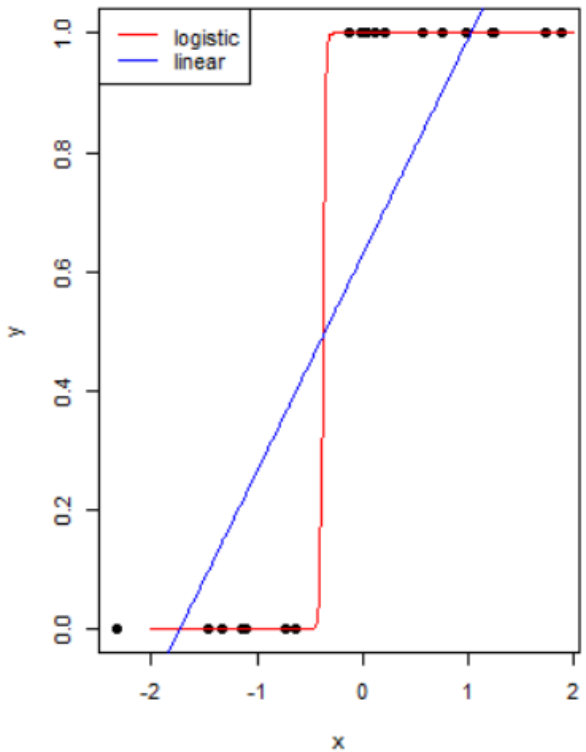
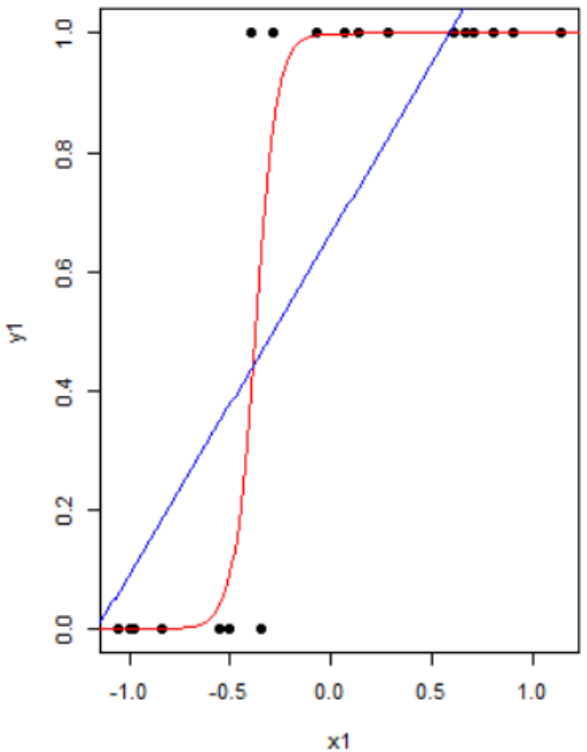
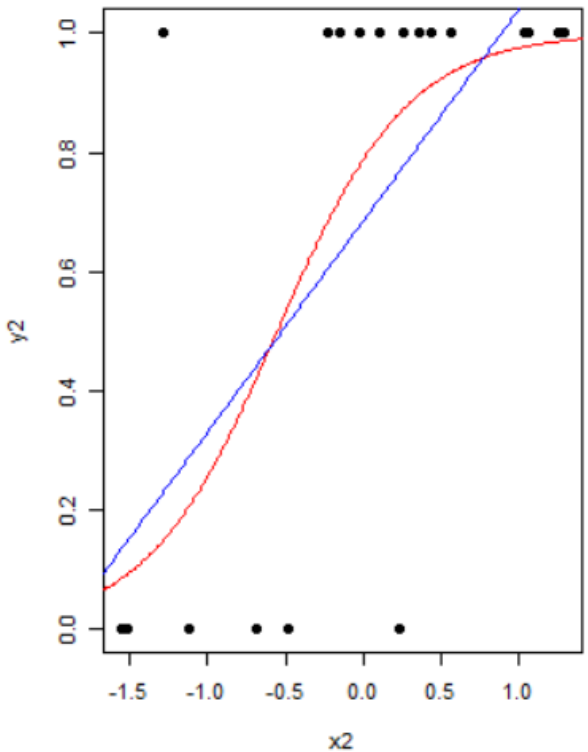


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Compared to the linear regression







Generalized linear model

The model could be generalized to be

$$Z_i = g(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

It means that the distribution of Y may be different from the normal distribution. It is defined by link functions:

$$g(\mu) = X\beta, (\mu = g^{-1}(X\beta)).$$

Distribution	Support of distribution	Typical uses	Link function, $\mathbf{X\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	$\mathbf{X\beta} = \mu$	$\mu = \mathbf{X\beta}$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	$\mathbf{X\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X\beta})^{-1}$
Gamma				
Inverse Gaussian	real: $(0, +\infty)$		$\mathbf{X\beta} = \mu^{-2}$	$\mu = (\mathbf{X\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	$\mathbf{X\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	$\mathbf{X\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X\beta})}{1+\exp(\mathbf{X\beta})} = \frac{1}{1+\exp(-\mathbf{X\beta})}$
Categorical	integer: $[0, K)$	outcome of single K-way occurrence	$\mathbf{X\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1			
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences		

Logistic regression

Rather than modeling the response Y directly, the logistic regression models "the probability that Y belongs to a particular category".
For the Default data, the probability of default given balance can be written as

$$\Pr(\text{default} = \text{Yes}| \text{ balance}).$$

Properties:

- The probability is between 0 and 1.
- For any given value of balance, a prediction can be made for the probability of default.
- One can predict "default $=$ Yes" if the probability > 0.5.
- If a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as the probability > 0.1.

Logistic regression

For convenience, we are using the generic 0/1 coding for the response.

$$Y = \begin{cases} 1, & \text{default} = \text{Yes} \\ 0, & \text{default} = \text{No.} \end{cases}$$

The probability model could be

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

$$\Pr(Y = 0|X) = 1 - \Pr(Y = 1|X) = 1 - \beta_0 - \beta_1 X,$$

or using logistic function is

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

$$\Pr(Y = 0|X) = 1 - \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} = \frac{1}{1 + e^{\beta_0+\beta_1 X}}.$$

Logistic regression

The logistic function is

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

and it implies the "odds" with

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

or the "log the odds" or "logit" is

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X.$$

Meaning in logistic regression

- Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively.
- Increasing X by one unit changes the log odds by β_1.
- Increasing X by one unit changes the odds by e^{β_1}.
- The fact that there is not a straight-line relationship between $p(X)$ and X.
- The rate of change in $p(X)$ per unit change in X depends on the current value of X.

Estimation (objective function)

- Normal assumption: Least square estimation
- non-normal assumption: maximum likelihood estimation

The likelihood function is

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{1_{\{y_i=1\}}} \left[1 - p(x_i)\right]^{1_{\{y_i=0\}}}.$$

The maximum likelihood estimates are the estimates that maximize $L(\beta_0, \beta_1)$.

Example: Default dataset (numerical results)

	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

TABLE 4.1. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance*. A one-unit increase in* balance *is associated with an increase in the log odds of* default *by* 0.0055 *units.*

- The estimate $\hat{\beta}_1 = 0.0055$, and it indicates that an increase in balance is associated with an increase in the probability of default.
- A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.
- The z-statistic associated with β_1 is equal to $\hat{\beta}_1/se(\hat{\beta}_1)$. A large (absolute) value of the z-statistic indicates evidence against the null hypothesis H_0: $\beta_1 = 0$.

Example: Default dataset (inference)

- The null hypothesis H_0 is

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

  which implies that that the probability of default does not depend on the changes of balance.
- Since the p-value is smaller than 0.05, then the null hypothesis is rejected.
- We conclude that there is indeed an association between balance and probability of default.
- The estimated intercept is typically not of interest. The main purpose is to adjust the average fitted probabilities to the proportion of ones in the data (in this case, the overall default rate).

Example: Default dataset (prediction)

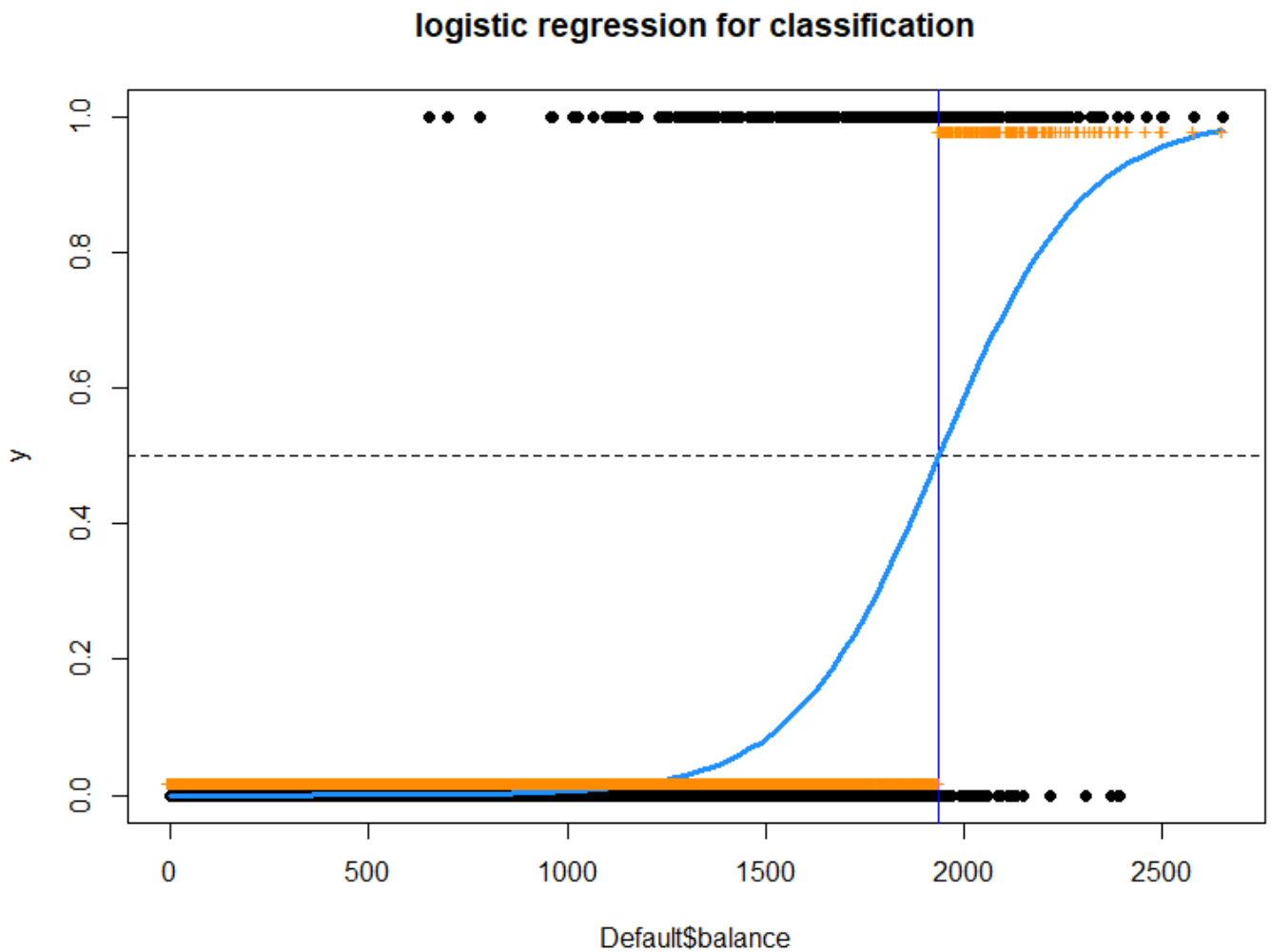
Once the coefficients have been estimated, we can compute the probability of default for any given credit card balance. Given balance $1,000$, the probability of default is

$$\hat{p}(1000) = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576.$$

Given balance $2,000$, the probability of default is

$$\hat{p}(2000) = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586.$$

Example: Default dataset (classification)



Example: Default dataset (qualitative predictor)

We can use qualitative predictors with the logistic regression model using the dummy variable approach.
Default data set contains the qualitative variable student, where student is YES or NO.

	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

TABLE 4.2. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using student status. Student status is encoded as a dummy variable, with a value of* 1 *for a student and a value of* 0 *for a non-student, and represented by the variable* `student[Yes]` *in the table.*

Question: What does the null hypothesis $H_0 : \beta_1 = 0$ mean?

Example: Default dataset (qualitative predictor)

- Dummy variables for both X and Y.
- Create a dummy variable that takes on a value of 1 for students and 0 for non-students.
- The coefficient associated with the dummy variable is positive, and the associated p-value is statistically significant.
- The probabilities of default for student and non-student groups are:

$$\hat{p}(\text{default=Yes}|\text{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431.$$

$$\hat{p}(\text{default=Yes}|\text{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Multiple logistic regression (multiple X's)

The multiple logistic regression with p covariates and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$.

$$g\left(E(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i)\right) = \log \frac{p_i(\boldsymbol{x}_i)}{1 - p_i(\boldsymbol{x}_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

which implies

$$p_i(\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i^T \beta}}{1 + e^{\boldsymbol{x}_i^T \beta}}.$$

Inference of logistic regression

Odds:

$$\frac{p}{1-p} = \exp\{\boldsymbol{x}^T\beta\}.$$

log-odds:

$$\ln \frac{p}{1-p} = \boldsymbol{x}^T\beta.$$

- If the value of $\boldsymbol{x}^T\beta$ is greater than 0, the probability of success increases.
- If the value of $\boldsymbol{x}^T\beta$ is smaller than 0, the probability of success decreases.

Estimation (objective function)

- Normal assumption: Least square estimation
- non-normal assumption: maximum likelihood estimation

The multiple logistic regression with p covariates and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$.

What is the method for estimating the parameters?

Logistic regression in classification

We use the value of $\boldsymbol{x}_i^T\beta$ as the judgement for classification.
Usually,

- If the value of $\boldsymbol{x}_i^T\beta$ is greater than and equal to 0, then classify the sample to the event of success.
- If the value of $\boldsymbol{x}_i^T\beta$ is smaller than 0, then classify the sample to the event of failure.

Example: Default dataset (multiple)

	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

TABLE 4.3. *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using* `balance`, `income`, *and student status. Student status is encoded as a dummy variable* `student[Yes]`, *with a value of* 1 *for a student and a value of* 0 *for a non-student. In fitting this model,* `income` *was measured in thousands of dollars.*

Question: What are the meaning of coefficients?

Example: Default dataset (multiple)

- The p-values associated with balance and the dummy variable for student status are very small.
- The coefficient for the dummy variable is negative, indicating that students are less likely to default than non-students. (why?)
- Confounding issue (see next page).
- The negative coefficient for student in the multiple logistic regression indicates that for "a fixed value of balance and income", a student is less likely to default than a non-student.
- The horizontal broken lines show the default rates for students and non-students averaged over all values of balance and income. The overall student default rate is higher than the non-student default rate.

Example: Default dataset (Confounding)

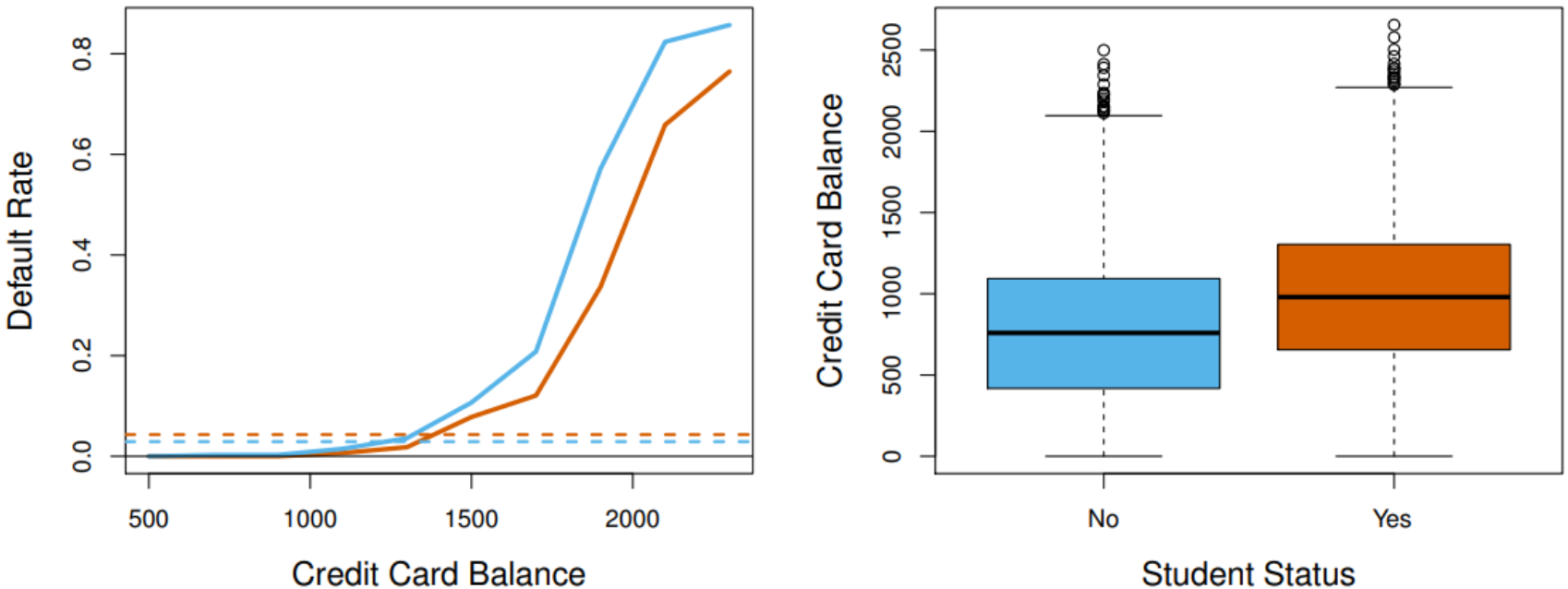


FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

Example: Default dataset (Confounding)

- The right-hand panel provides an explanation that the variables student and balance are correlated.
- Students tend to hold higher levels of debt, which is in turn associated with higher probability of default.
- Students are more likely to have large credit card balances, which tend to be associated with high default rates.
- The phenomenon seen in Figure 4.3 is known as confounding.
- The student is less risky than a non-student with the same credit card balance.

Multinomial logistic regression (multiple levels of Y)

Sometimes, we classify a response variable (Y) that has more than two classes. Assume that there are K classes for Y, defined by $Y = 1, \ldots, k$.

- We first select a single multinomial logistic regression class to serve as the baseline.
- Without loss of generality, we select the Kth class for the baseline.
- The multinomial multiple logistic regression with p covariates and

$$\log \frac{p(Y = j|\boldsymbol{X} = \boldsymbol{x})}{1 - p(Y = j|\boldsymbol{X} = \boldsymbol{x})} = \beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p \; j = 1, \ldots, K - 1.$$

- The log odds between any pair of classes is linear in the features.
- The probability of the jth class is

$$p(Y = j|\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{jp}x_p}}{1 + \sum_{j=1}^{k-1} e^{\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{jp}x_p}}.$$

Generalized linear model

The model could be generalized to be

$$Z_i = g(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

It means that the distribution of Y may be different from the normal distribution. It is defined by link functions:

$$g(\mu) = X\beta, (\mu = g^{-1}(X\beta)).$$

Distribution	Support of distribution	Typical uses	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Categorical	integer: $[0, K)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences		

Further models

Statistical Model: $Y = f(x_1, x_2, \ldots, x_p) + \epsilon$.

- $f(x)$
  1. linear form (linear function of unknown parameters)

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2,$$

$$E(y) = \beta_0 + \beta_1 x_1 + exp(\beta_2) x_2,$$

  2. nonlinear form: $E(y) = \exp\{\theta_1 x_1 \exp(-\theta_2 x_2)\}$
  3. categorical variables
- distribution of ϵ
  1. $N(0, \sigma^2)$: regression
  2. non-normal distribution: generalized model, logistic regression, probit model, ...