CLASSIFICATION

Part I

Outline

- An overview of classification
- Logistic regression

An overview of classification

Classification

- The task of predicting a qualitative or categorical response
  - E.g., disease status

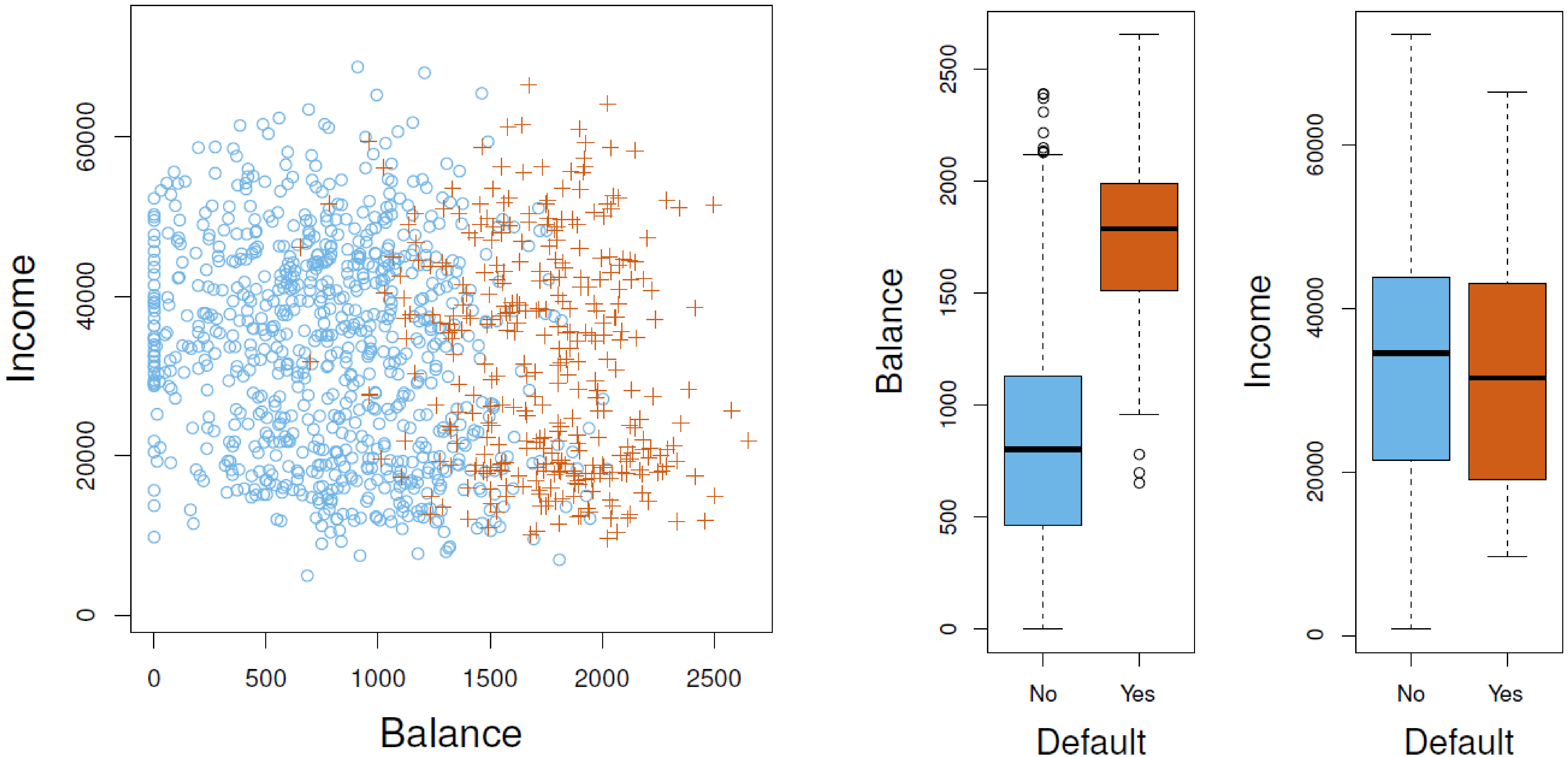
Examples

- *A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?*
- *An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth*

- Popular classification methods, or *classifiers*, include
  - K-nearest neighbors
  - Logistic regression
  - Linear discriminant analysis
  - Naive Bayes

Default data

- Simulated customer default records for a credit card company
- The goal is to predict whether an individual will default on his or her credit card payment, on the basis of annual income, monthly credit card balance, and other factors

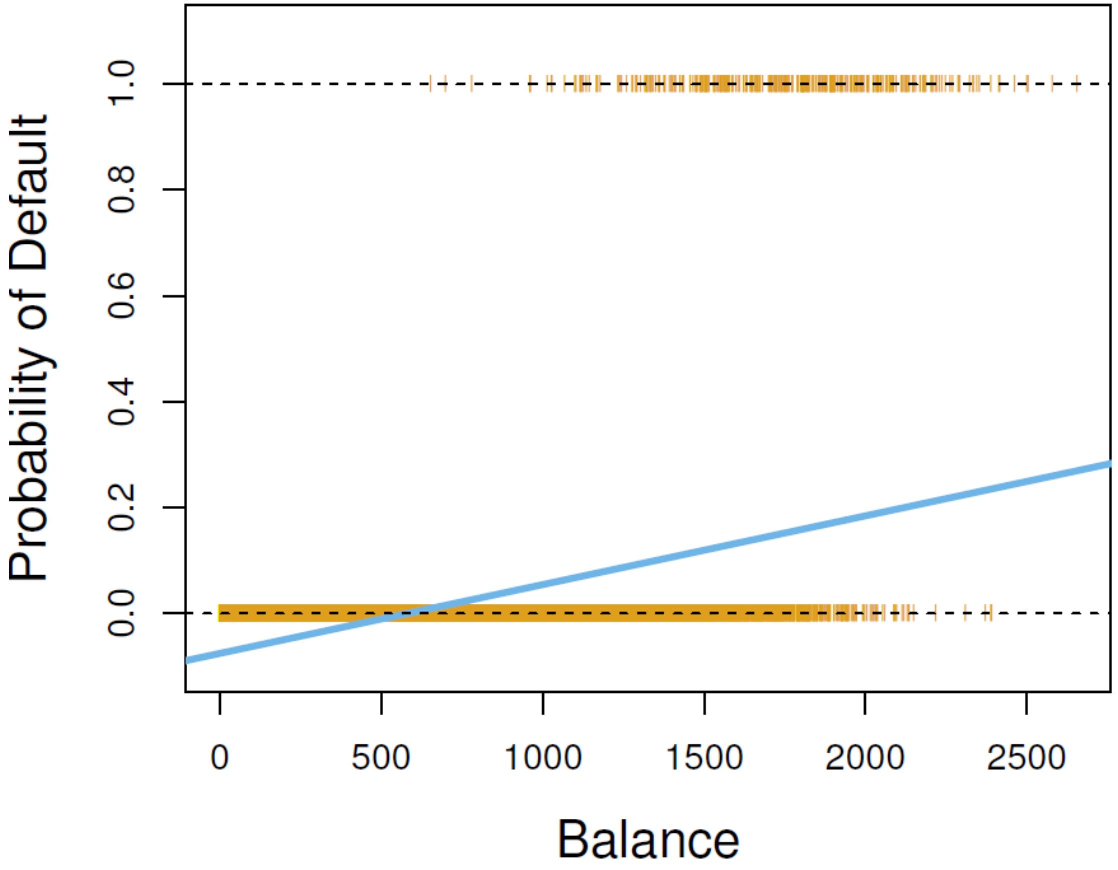


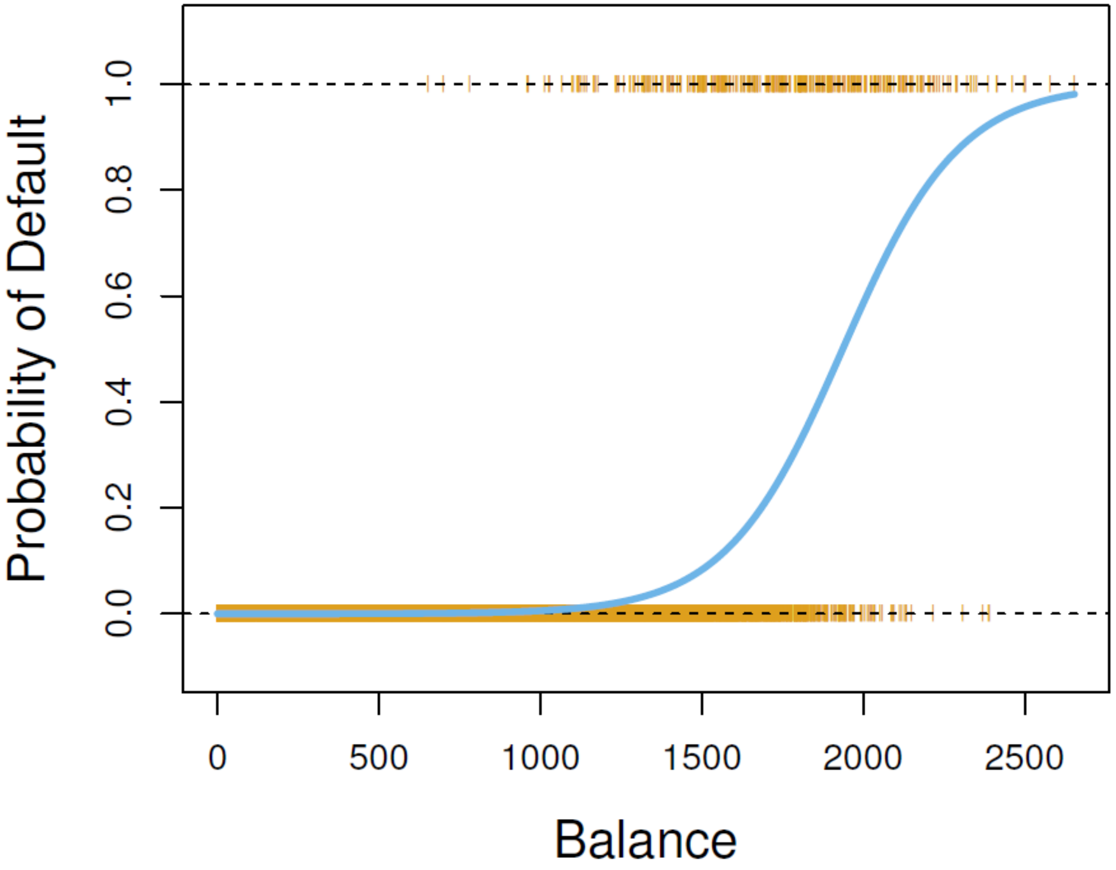
Some of the figures and tables in this presentation are taken from "*An Introduction to Statistical Learning, with Applications in R*" (Springer) with permission from the authors: G. James, D. Witten, T. Hastie, and R. Tibshirani

Why not linear regression?

- Convert a qualitative response into a quantitative response

- Binary responses
  - The dummy variable approach
- Responses with more than two levels
  - E.g., stroke, drug overdose, and epileptic seizure





Some of our estimates might be outside the $[0, 1]$ interval

Logistic regression

- Modeling the *conditional* probability that the response belongs to a particular category, given the observed predictors
  - E.g., the probability of default given balance

Binary responses

- Use the 0/1 coding scheme
- Define $p(X) = \Pr(Y = 1|X)$

- In logistic regression

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

or

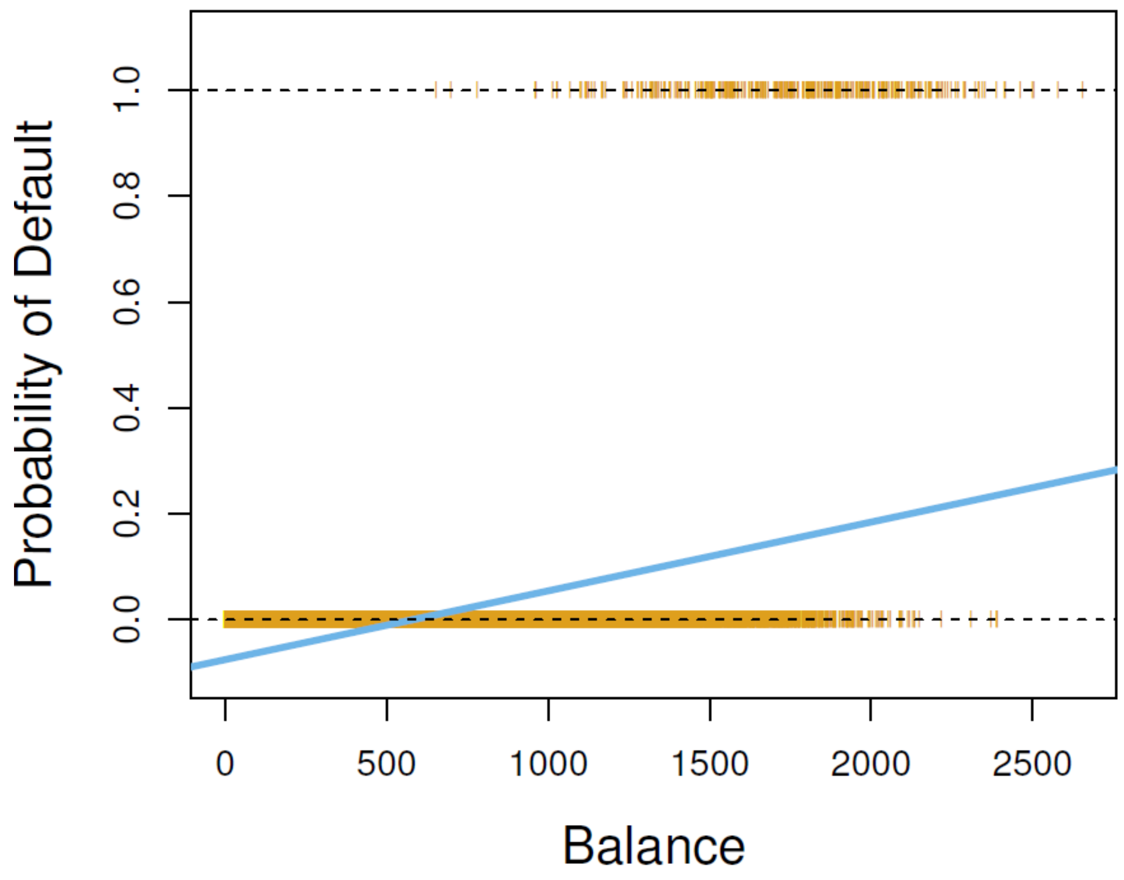
$$\log\left\{\frac{p(X)}{1-p(X)}\right\} = \beta_0 + \beta_1 X$$

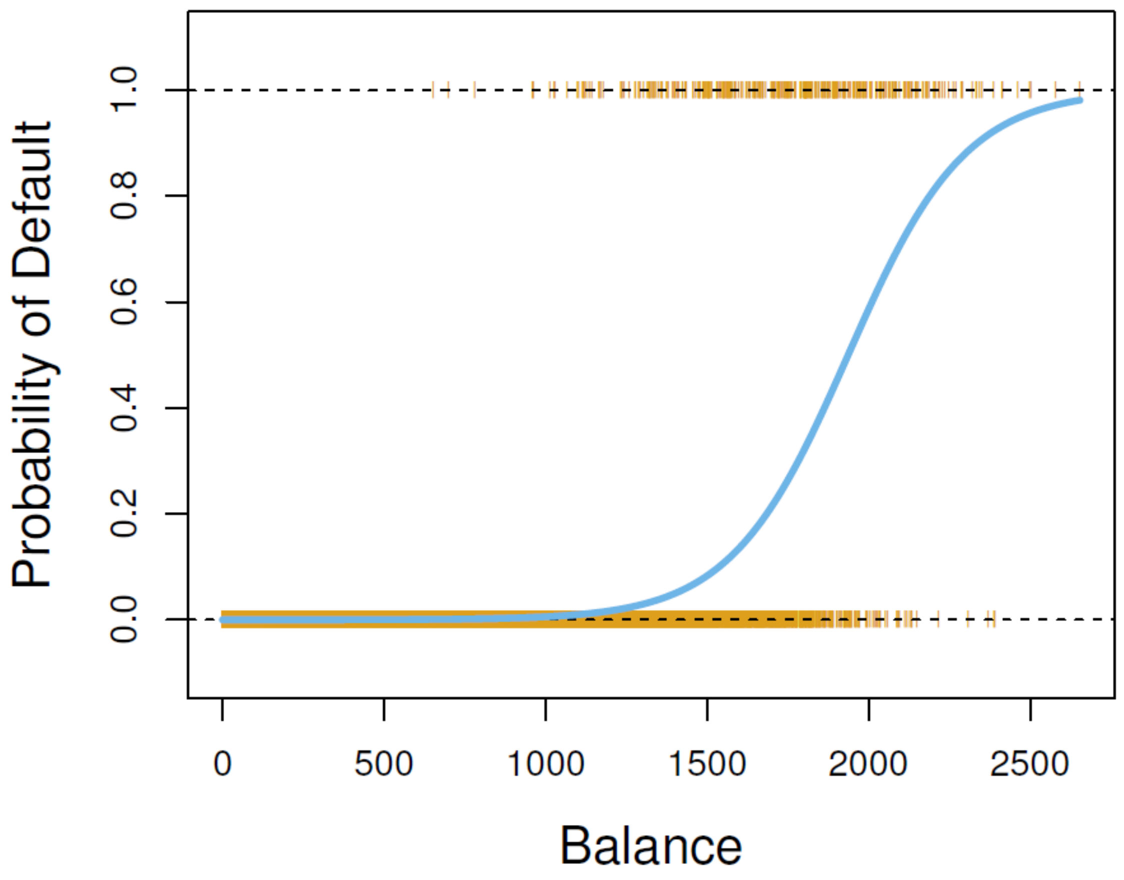
- $p(X)/\{1 - p(X)\}$ is called the *odds*
- The log of odds is called the *logit*
  - Increasing X by one unit changes the logit by β_1, or equivalently it multiplies the odds by e^{β_1}

- In linear regression

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

  - β_1 gives the average change in Y associated with a one-unit increase in X


Probability of Default
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Balance


Probability of Default
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Balance

Maximum likelihood estimates

- The likelihood function

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

- Maximum likelihood chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to maximize the likelihood function

- The estimated probability

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

- The prediction rule?

	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.*

	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.*

Multiple logistic regression

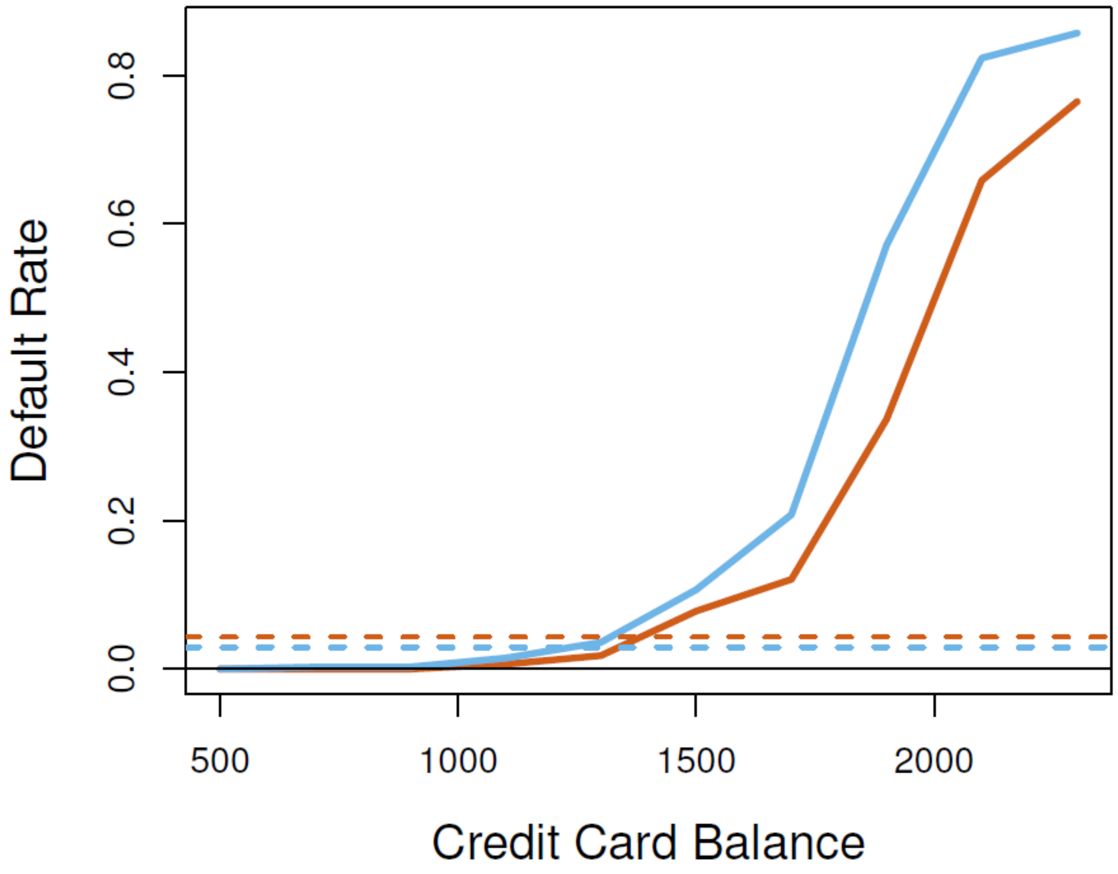
$$\frac{p(X_1, X_2, \dots, X_p)}{1 - p(X_1, X_2, \dots, X_p)} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$$

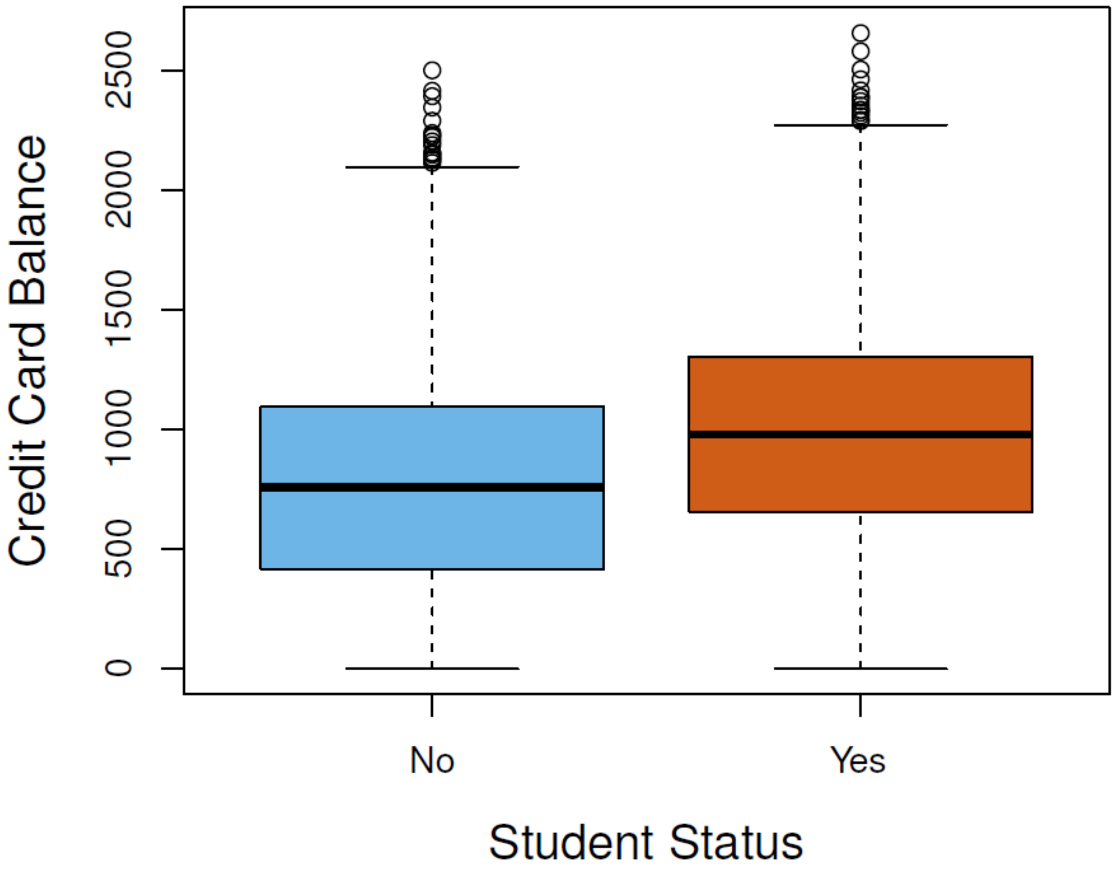
or

$$\log\left\{\frac{p(X_1, X_2, \dots, X_p)}{1 - p(X_1, X_2, \dots, X_p)}\right\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

	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.*





Confounding in the Default data

More than two response classes?

- Extend the two-class logistic regression approach to the setting of $K > 2$ classes

Multinomial logistic regression

- Without loss of generality, select the Kth class to serve as the baseline
- Assume that

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

for $k = 1, \ldots, K-1$, and

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

- Show that

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

  - The log odds between any pair of classes is linear in the features
- The decision to treat the Kth class as the baseline is unimportant
  - Interpretation of the coefficients is tied to the choice of baseline and must be done with care

The softmax coding

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

or

$$\log\left\{\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right\} = \beta_{kk'0} + \beta_{kk'1}x_1 + \cdots + \beta_{kk'p}x_p$$