Lecture 7: Classification (Textbook 4.1-4.3)

Classification

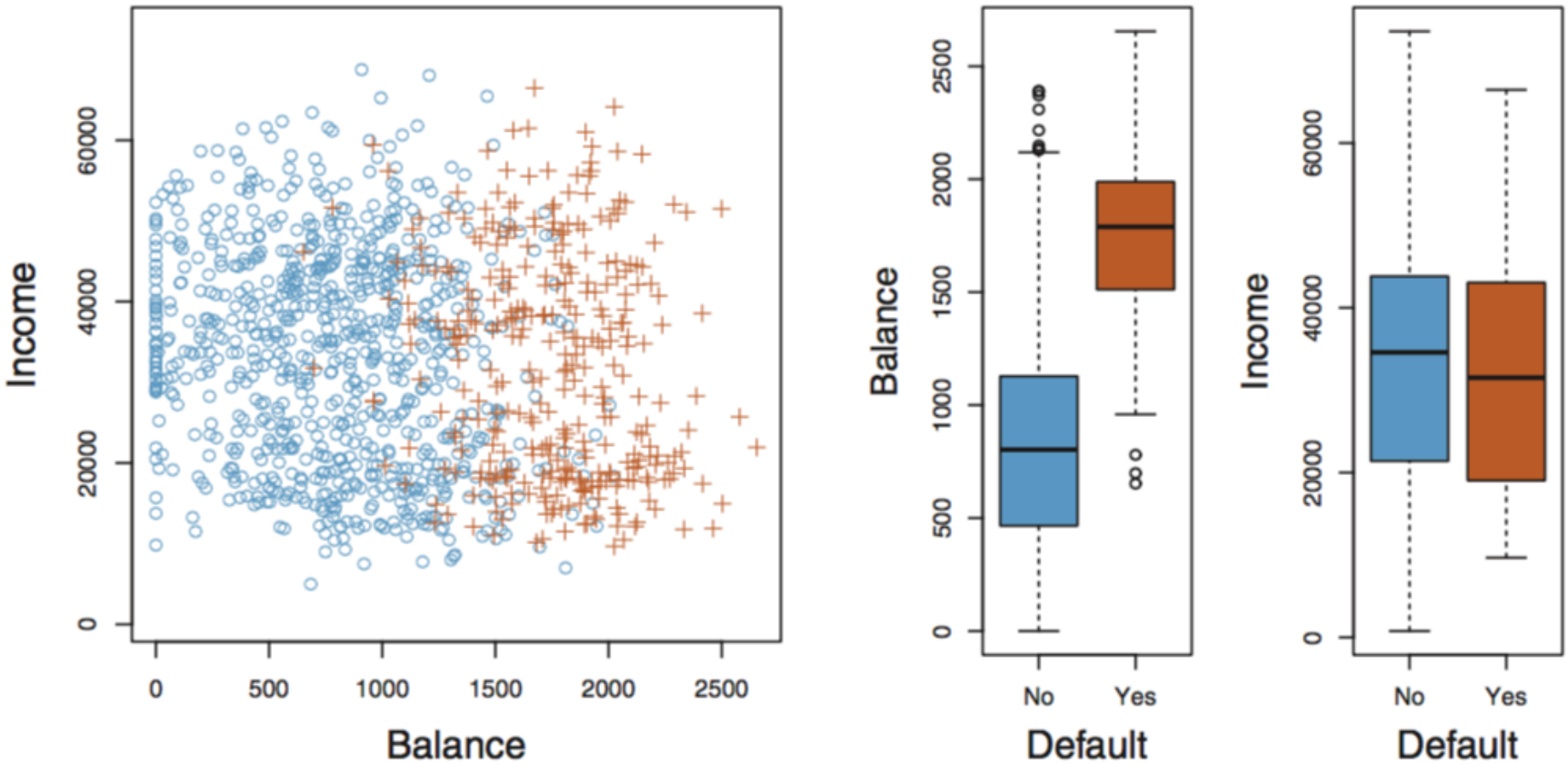
- Qualitative variables take values in an unordered set C, such as:

  eye color$\in$ {brown,blue,green},
  email$\in$ {spam, ham}.

- Our goal: Given a feature vector X and a qualitative response Y taking values in the set C, we aim to build a function $C(X)$ that uses the feature vector X to predict Y; i.e. $C(X) \in C$.

- In this chapter we discuss three of the most widely-used classifiers: **logistic regression**, **linear discriminant analysis**, and **K-nearest neighbors**.

Default Data

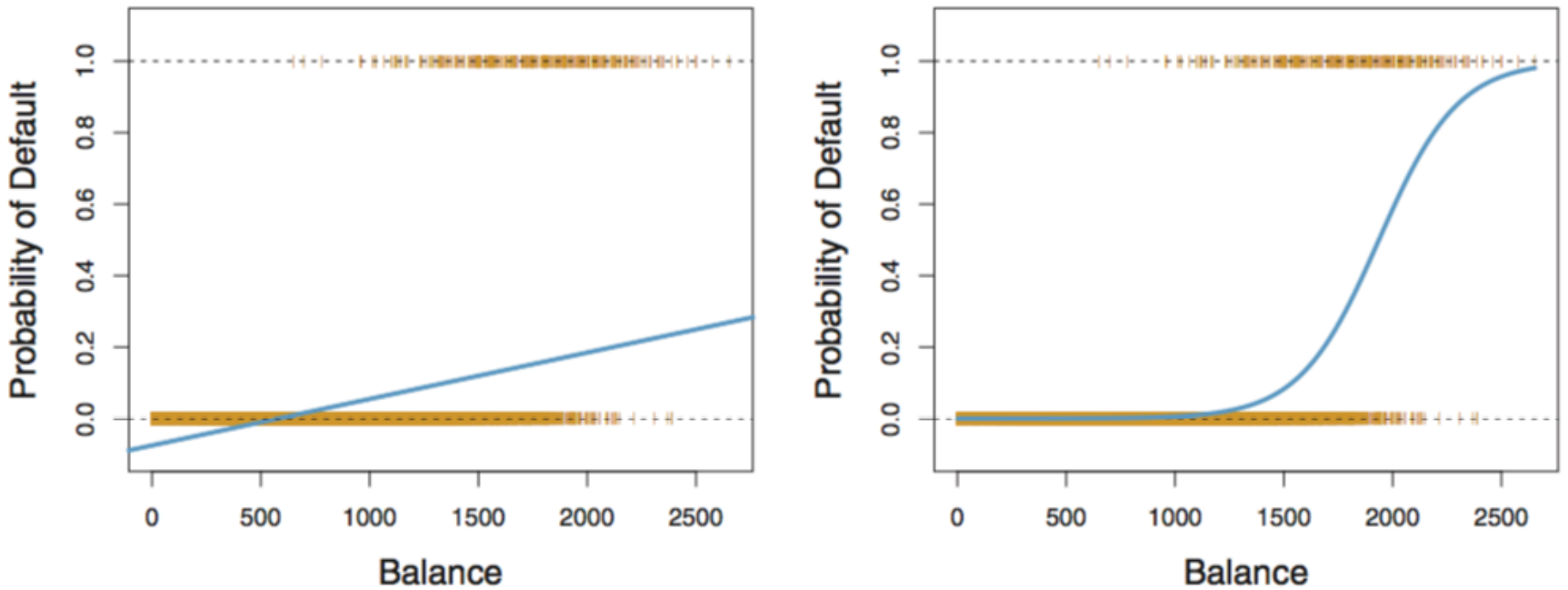


It shows the annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.

Why Not Linear Regression?

- In this case of a binary outcome, linear discriminant analysis, which is related but different from linear regression, does a good job as a classifier.
- The least squared method can estimate $E(Y|X = x) = Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- However, linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.

Linear versus Logistic Regression



Left: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange points represents 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.

Logistic Regression

- How to model the relationship between $p(X) = Pr(Y = 1|X)$ and X?
- A linear regression may estimate $p(X) < 0$ or $p(X) > 1$.
- Logistic regression model $p(X)$ by the logistic function,

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

  It is easy to see that no matter what values β_0, β_1, or X take, $p(X)$ will have values between 0 and 1.

- Note that $p(X)$ is Not a linear function X or β.

Logistic Regression

- A bit of rearrangement gives

$$\underbrace{\frac{p(X)}{1-p(X)}}_{odds} = e^{\beta_0+\beta_1 X}, \quad \underbrace{\log\left[\frac{p(X)}{1-p(X)}\right]}_{log-odds} = \beta_0 + \beta_1 X.$$

  The odds takes value between 0 and $+\infty$, and the log odds takes value between $-\infty$ and $+\infty$.

- β_1 represents the change of log odds by increasing X by one unit, since

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

Maximum Likelihood

Given training data $(x_1, y_1), ..., (x_n, y_n)$, we use **maximum likelihood** to estimate the parameters.

The maximum likelihood principle is that we seek the estimates of parameters such that the fitted probability corresponds as closely as possible to the individual's observed outcome.

The **likelihood function** of the observed data is

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

We pick β_0 and β_1 to maximize the likelihood of the observed data.

Most statistical packages can fit logistic regression models by maximum likelihood. In R we use the glm function.

Inference

Consider again the Default data (fitted by maximum likelihood):

	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

Z-statistic is similar to t-statistic in regression, and is defined as

$$\hat{\beta}_1 / SE(\hat{\beta}_1).$$

It produces p-value for testing the null hypothesis $H_0 : \beta_1 = 0$. A large (absolute) value of the z-statistic or small p-value indicates evidence against H_0.

Making Predictions

Consider the Default data with student as predictor. What is our estimated probability of default for a student? To fit the model, we create a dummy variable that takes on a value of 1 for students and 0 for non-students.

	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

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

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

Multiple Logistic Regression

Consider the Default data using balance, income, and student status as predictors.

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

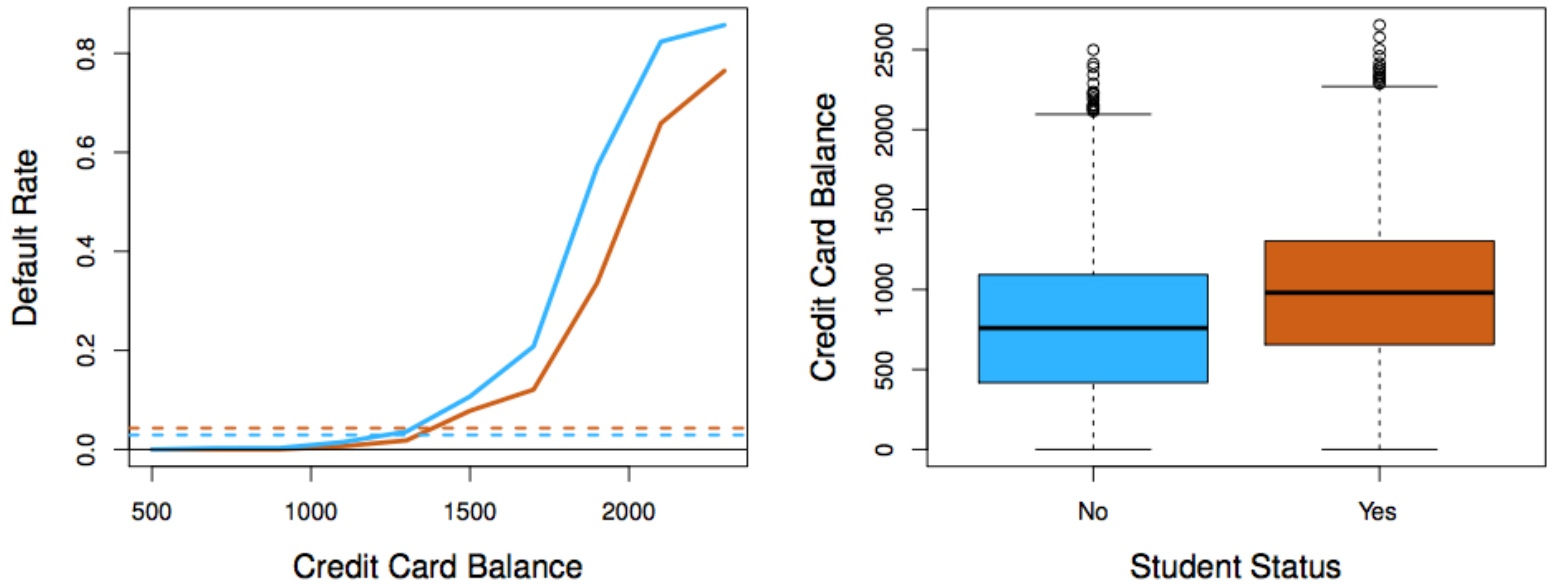
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	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

Why is coefficient for student negative, while it was positive before?

Confounding

The results obtained using one predictor may be quite different from those obtained using multiple predictors, especially when there is correlation among the predictors. The phenomenon is known as **confounding**.



- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.