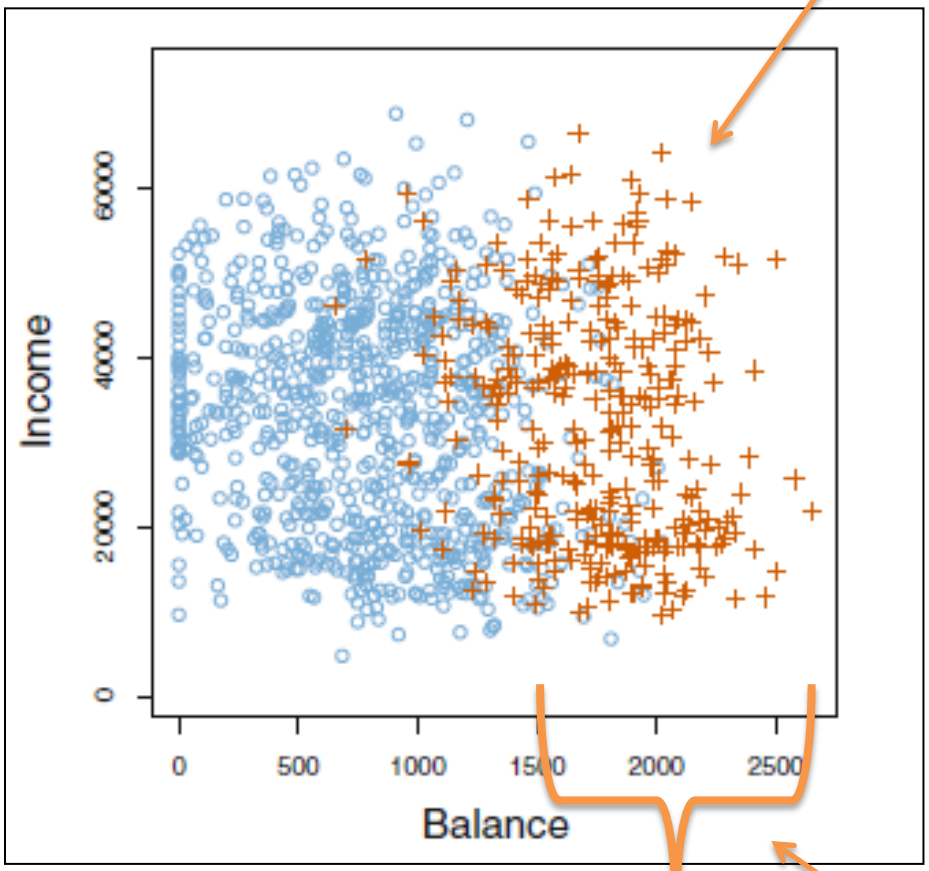
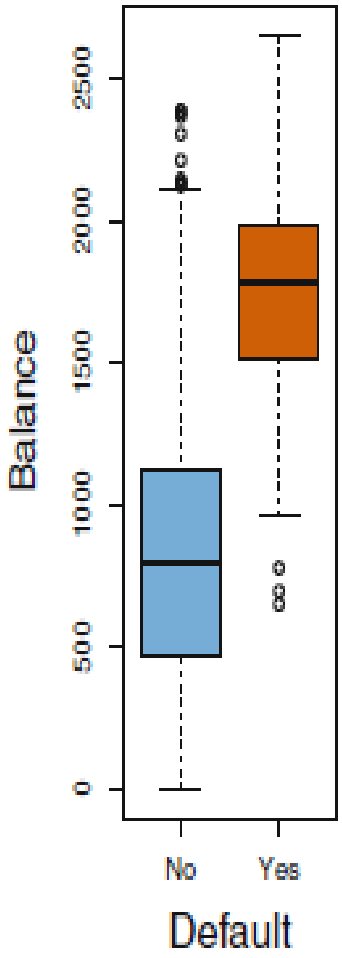
- Variável resposta qualitativa (categórica).
- Métodos para predizer respostas qualitativas: Classificação.
  - Regressão Logística
  - Análise de “descriminante linear”
  - K-nearest neighbours
- Problemas de Classificação – Exemplos:
  - A) classificar uma situação médica de uma pessoa com determinados sintomas em 3 possíveis diagnósticos.
  - B) Determinar se uma transação é fraudulenta se baseando em características como endereço IP, transações anteriores, por exemplo.
  - C) classificar a partir do sequenciamento de DNA quais genes são responsáveis por determinada doença.
- Temos aqui também um conjunto de treinamento que será usado para construir um classificador.
- Neste capítulo, o conceito de classificação será ilustrado usando um conjunto de dados simulado sobre inadimplência (default) em função da renda anual (incomes) e saldo mensal do cartão de crédito (monthly credit card balance).

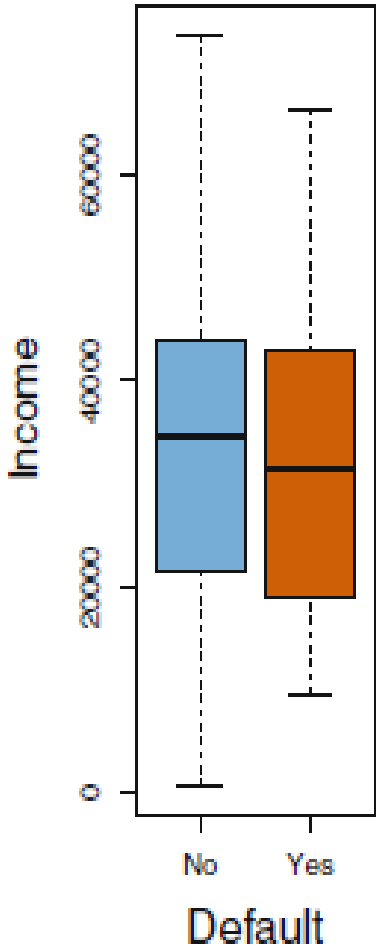
10.000 indivíduos

Pessoas que não pagaram a fatura em algum mês.



Quem deixa de pagar a fatura de cartão de crédito tem saldos mais altos.





Quanto maior o saldo da fatura maior é a chance de inadimplência.

Vamos estudar a

- Variável resposta Y (inadimplência) em função
- Do saldo (X_1)
- E da renda (X_2)

Nosso principal objetivo é predizer “não pagamento”, para quaisquer valores de renda anual e saldo mensal (característica que envolve quanto a pessoa gasta mensalmente.)

Y é uma variável categórica que assume “Yes” ou “No”.

A regressão logística responde à pergunta: Qual é a probabilidade de Y assumir Yes ou No ?

$P(\{Y = Yes\} \mid X\_1)$, onde X_1: saldo (balance)

Uma vez que tenhamos essa distribuição de probabilidade condicional, podemos escolher critérios (pontos de corte) para classificar os usuários de cartão de crédito.

No texto o exemplo $P(\{Y = Yes\} \mid X\_1) > 0.5$ para uma instituição e $P(\{Y = Yes\} \mid X\_1) > 0.1$ para outra instituição, bem mais conservadora.

$$p(X) = \Pr(Y = 1|X)$$

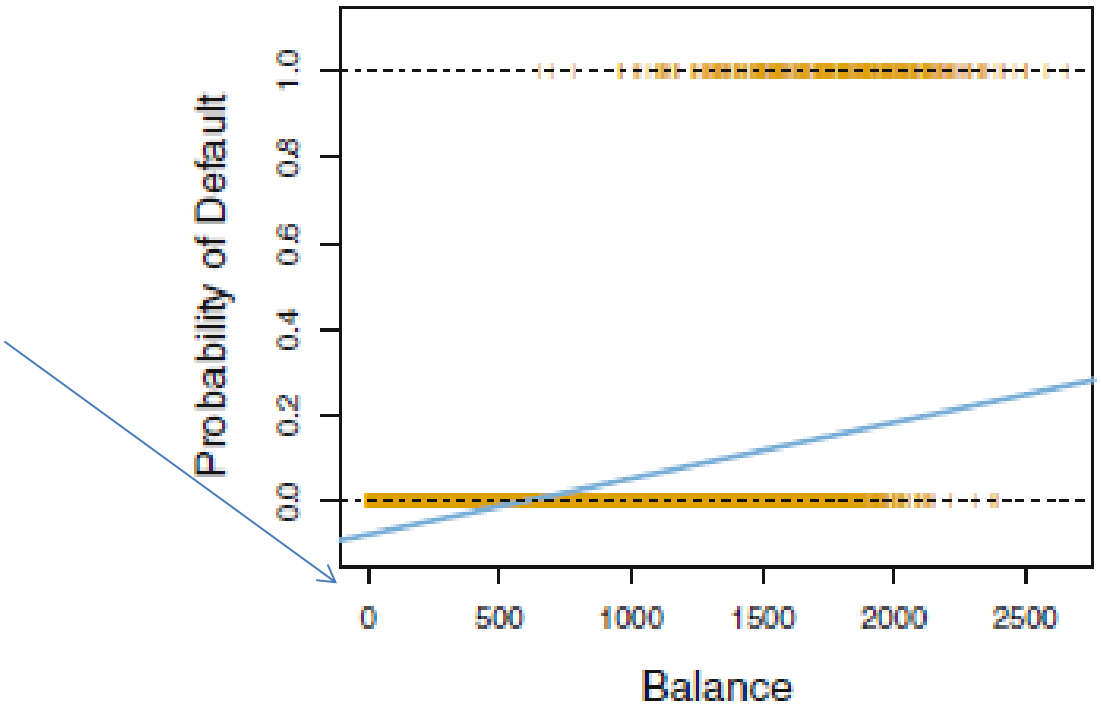
Notação do texto, para qualquer X, e Y binária ou categórica.

Se “Yes” Y=1
Se “No” Y=0

Será que podemos usar o modelo de regressão linear?

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

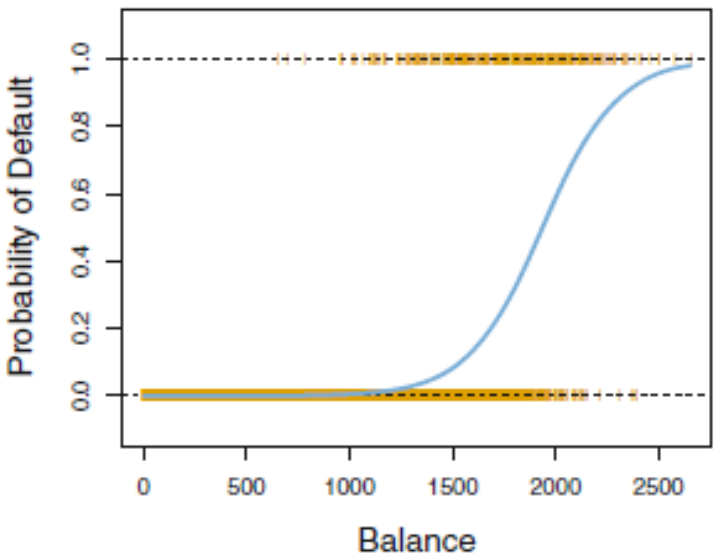
Valores negativos



Toda vez que tentarmos ajustar uma reta a um problema com variável resposta binária (0/1) teremos problemas com p(X).

Temos que modelar p(X) deforma a assumir valores no intervalo real [0,1]. No caso da regressão logística é usada a "função logística".

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



Não temos o problema da p(X) assumir valores negativos ou maiores que 1.

Comparação

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

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

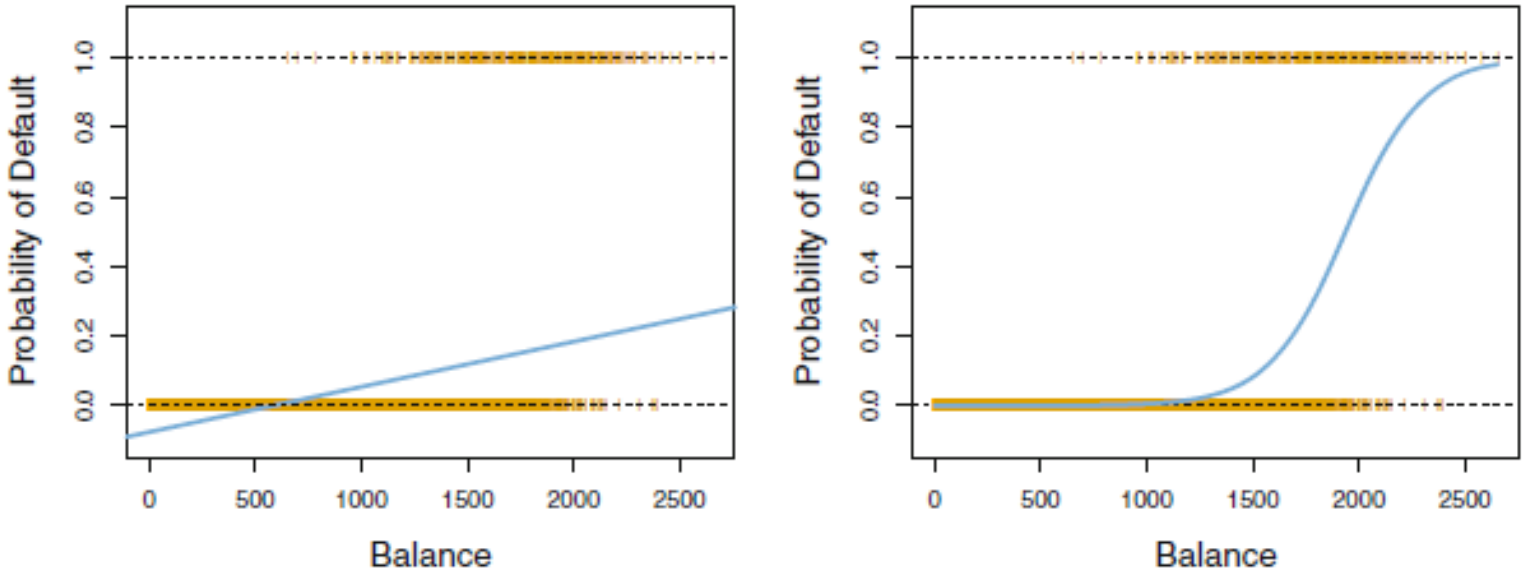


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using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

	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.*

Saldo de 1000 dólares:

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

Menos de 1% da chance.
Para $ 2000 a chance é 58.6 %.