Cross-validation

STAT 471

October 5, 2021

Where we are

- **Unit 1:** Intro to modern data mining
- **Unit 2:** Tuning predictive models
- **Unit 3:** Regression-based methods
- **Unit 4:** Tree-based methods
- **Unit 5:** Deep learning

Lecture 1: Logistic regression

Lecture 2: Regression in high dimensions

Lecture 3: Ridge regression

[**Fall break:** No class]

Lecture 4: Lasso regression

Lecture 5: Unit review and quiz in class

Homework 1 due the following Sunday.

Midterm exam following Monday (7-9pm).

Setting: Binary classification

```
> Default
# A tibble: 10,000 x 4
   default student balance income
   <fct>   <fct>     <dbl>  <dbl>
 1 No      No         730. 44362.
 2 No      Yes        817. 12106.
 3 No      No        1074. 31767.
 4 No      No         529. 35704.
 5 No      No         786. 38463.
 6 No      Yes        920.  7492.
 7 No      No         826. 24905.
 8 No      Yes        809. 17600.
 9 No      No        1161. 37469.
10 No      No           0  29275.
# … with 9,990 more rows
```

Will a person default on their credit card bill?

We build a model to approximate

$$\mathbb{P}(\text{default} = \text{Yes} \,|\, \text{student, balance, income})$$

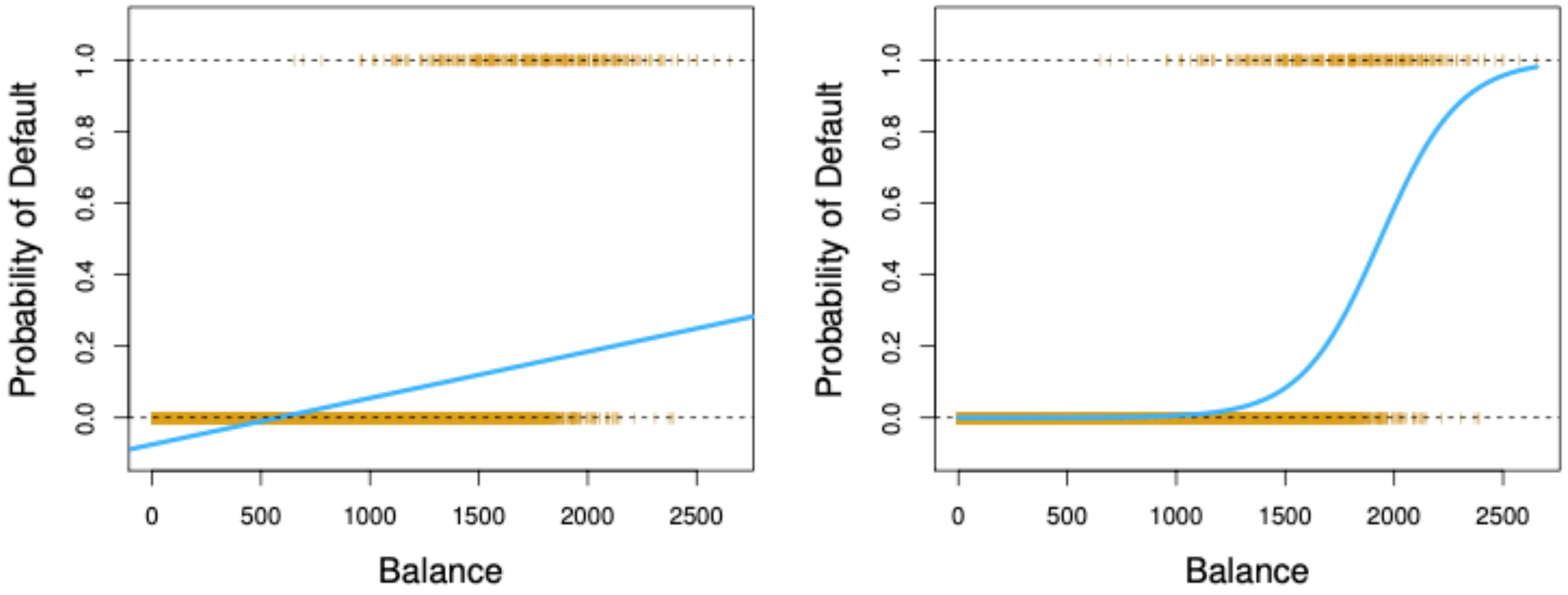
and then predict

$$\text{default} = \begin{cases} \text{Yes}, & \text{if } \widehat{\mathbb{P}}[\text{default}] \geq c; \\ \text{No}, & \text{if } \widehat{\mathbb{P}}[\text{default}] < c\,. \end{cases}$$

Most common choice: logistic regression model.

Why not linear regression?

$$\mathbb{P}[\text{default} \,|\, \text{balance}] = \beta_0 + \beta_1 \cdot \text{balance}$$

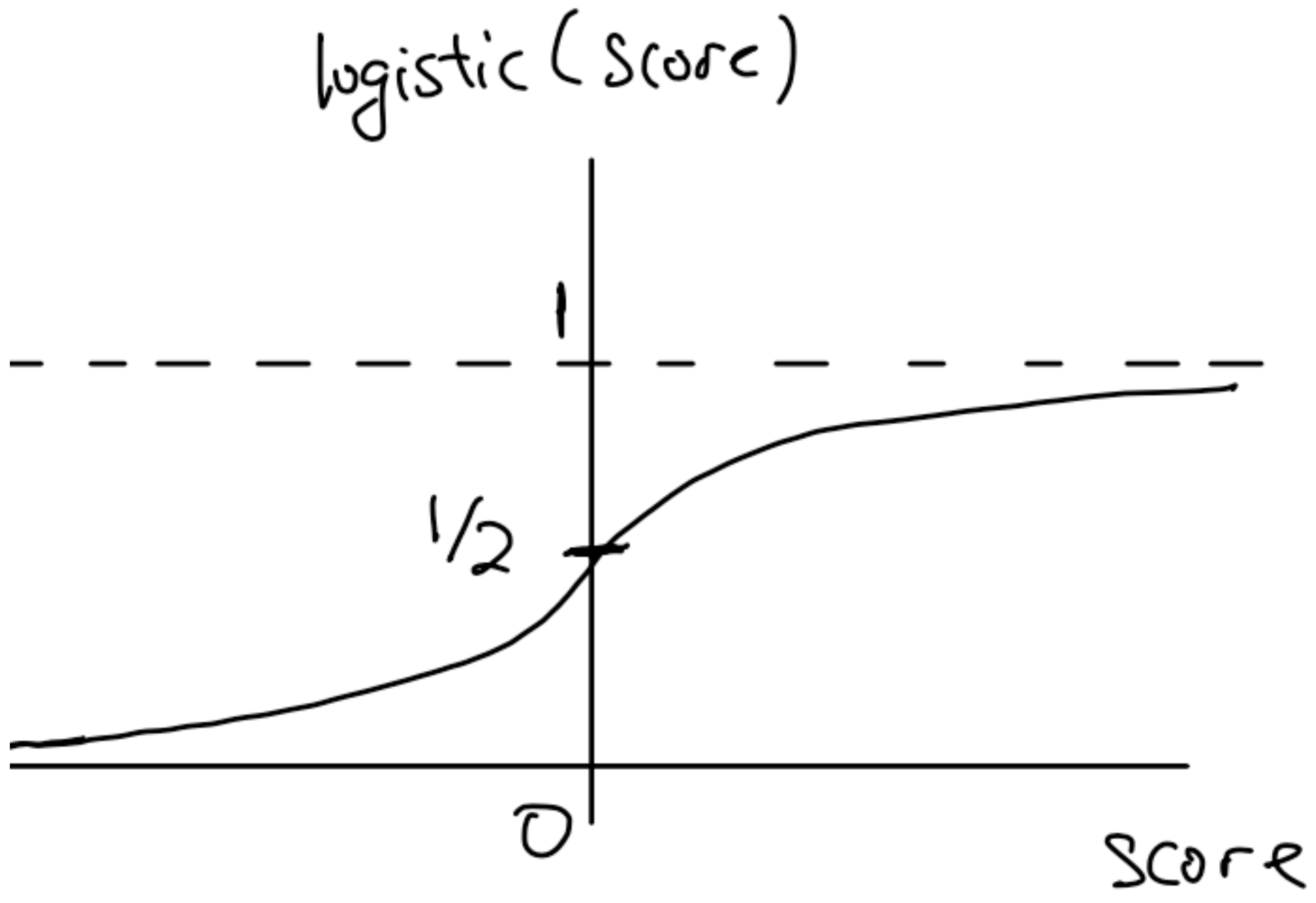


How do we get that nice smooth curve on the right?

The logistic transformation

Idea: use $\beta_0 + \beta_1 \cdot \text{balance}$ as a "score", then map the score onto [0,1] using a transformation!

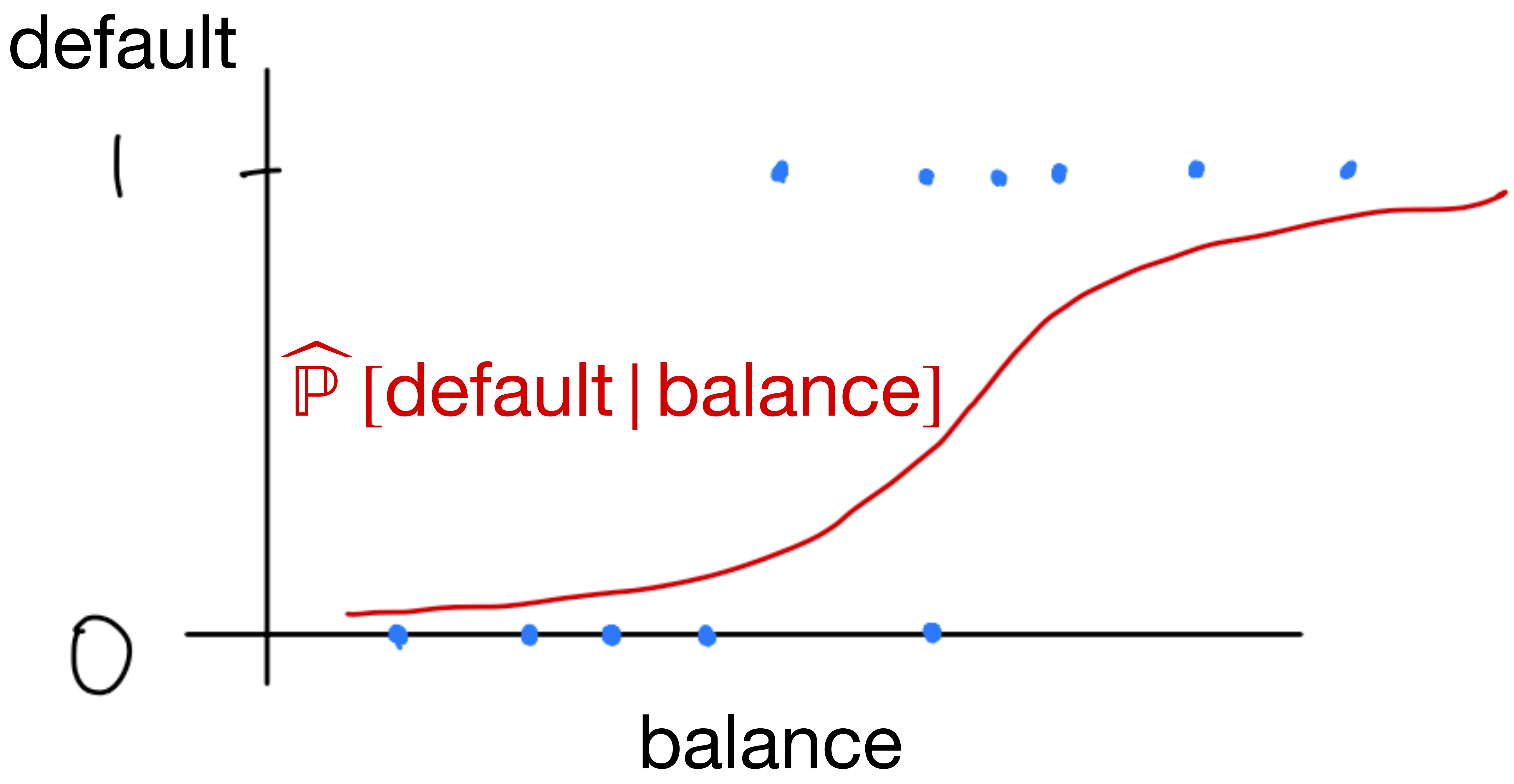
$$\text{logistic}(\text{score}) = \frac{e^{\text{score}}}{1 + e^{\text{score}}}$$



The logistic regression model

$$\cancel{\mathbb{P}[\text{default} \mid \text{balance}] = \beta_0 + \beta_1 \cdot \text{balance}}$$

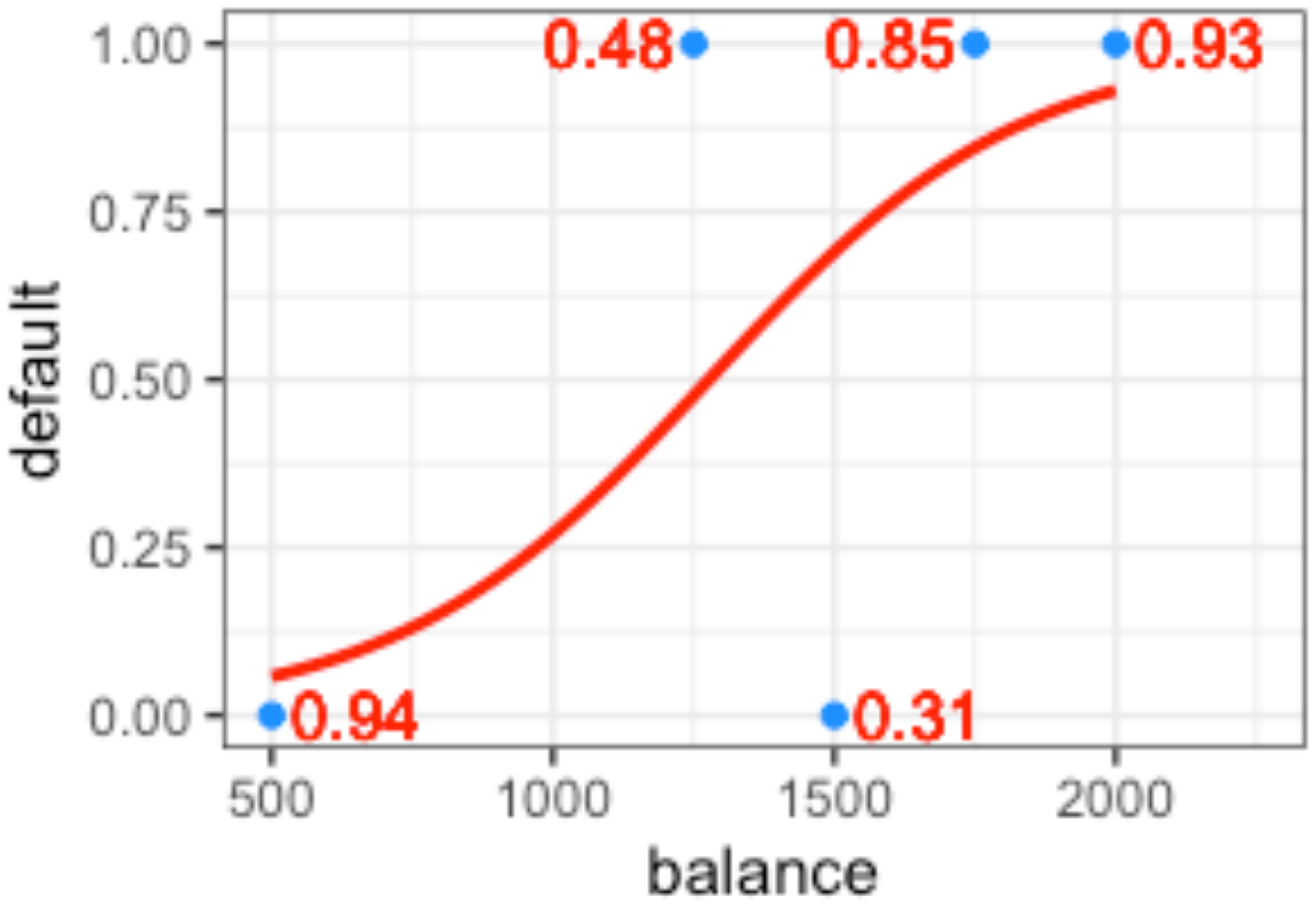
$$\mathbb{P}[\text{default} \mid \text{balance}] = \text{logistic}(\beta_0 + \beta_1 \cdot \text{balance})$$



Maximum likelihood estimation

Given (β_0, β_1), the probability $\mathscr{L}(\beta_0, \beta_1)$ of the observed data is called the likelihood.

Choose $(\widehat{\beta}_0, \widehat{\beta}_1)$ to maximize the likelihood.



Toy data

default	balance	P[default = 1]	P[observed]
1	\$1250	$\frac{e^{\beta_0+\beta_1\cdot 1250}}{1+e^{\beta_0+\beta_1\cdot 1250}}$	$\frac{e^{\beta_0+\beta_1\cdot 1250}}{1+e^{\beta_0+\beta_1\cdot 1250}}$
0	\$500	$\frac{e^{\beta_0+\beta_1\cdot 500}}{1+e^{\beta_0+\beta_1\cdot 500}}$	$\frac{1}{1+e^{\beta_0+\beta_1\cdot 500}}$
1	\$2000	$\frac{e^{\beta_0+\beta_1\cdot 2000}}{1+e^{\beta_0+\beta_1\cdot 2000}}$	$\frac{e^{\beta_0+\beta_1\cdot 2000}}{1+e^{\beta_0+\beta_1\cdot 2000}}$
1	\$1750	$\frac{e^{\beta_0+\beta_1\cdot 1750}}{1+e^{\beta_0+\beta_1\cdot 1750}}$	$\frac{e^{\beta_0+\beta_1\cdot 1750}}{1+e^{\beta_0+\beta_1\cdot 1750}}$
0	\$1500	$\frac{e^{\beta_0+\beta_1\cdot 1500}}{1+e^{\beta_0+\beta_1\cdot 1500}}$	$\frac{1}{1+e^{\beta_0+\beta_1\cdot 1500}}$

$$\mathscr{L}(\beta_0, \beta_1) = \frac{e^{\beta_0+\beta_1\cdot 1250}}{1+e^{\beta_0+\beta_1\cdot 1250}} \times \frac{1}{1+e^{\beta_0+\beta_1\cdot 500}} \times \frac{e^{\beta_0+\beta_1\cdot 2000}}{1+e^{\beta_0+\beta_1\cdot 2000}} \times \frac{e^{\beta_0+\beta_1\cdot 1750}}{1+e^{\beta_0+\beta_1\cdot 1750}} \times \frac{1}{1+e^{\beta_0+\beta_1\cdot 1500}}$$

Multiple logistic regression

Like with linear regression, can include multiple features, e.g.

$$\mathbb{P}[\text{default} \mid \text{student}, \text{balance}, \text{income}]$$
$$= \text{logistic}(\beta_0 + \beta_1 \cdot \text{student} + \beta_2 \cdot \text{balance} + \beta_3 \cdot \text{income})$$

The logistic regression likelihood, as well as the maximum likelihood estimates $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)$ are defined analogously.

Interpreting logistic regression coefficients

$$\mathbb{P}[\text{default}] = \text{logistic}(\beta_0 + \beta_1 \cdot \text{student} + \beta_2 \cdot \text{balance} + \beta_3 \cdot \text{income})$$

$$\Downarrow$$

$$\log \frac{\mathbb{P}[\text{default}]}{1 - \mathbb{P}[\text{default}]} = \beta_0 + \beta_1 \cdot \text{student} + \beta_2 \cdot \text{balance} + \beta_3 \cdot \text{income}$$

log-odds

$$\Downarrow$$

Increasing balance by 500 while controlling for the other features tends to (additively) increase the log-odds of default by $500 \cdot \beta_2$.

$$\Downarrow$$

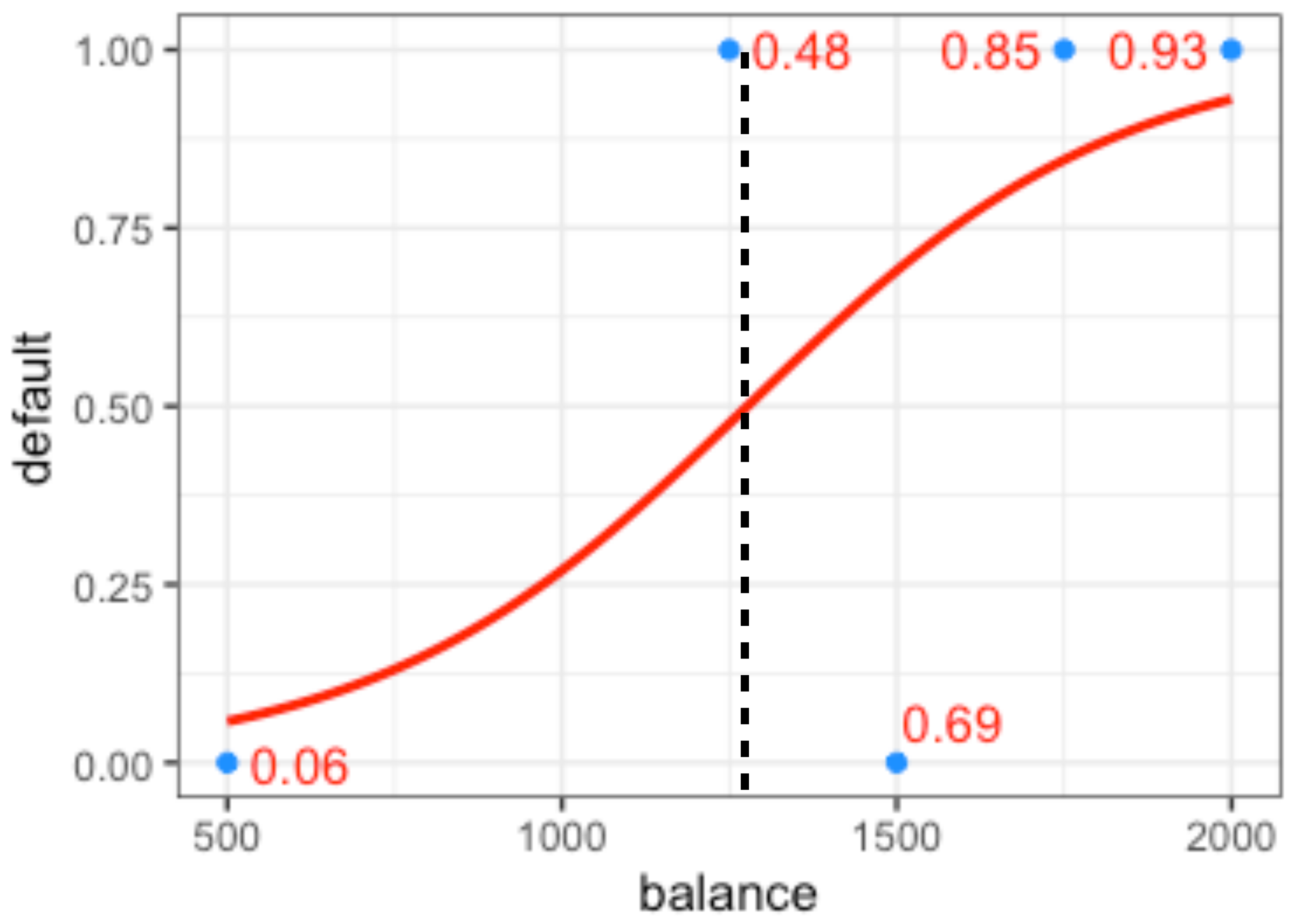
Increasing balance by 500 while controlling for the other features tends to (multiplicatively) increase the odds of default by $e^{500 \cdot \beta_2}$.

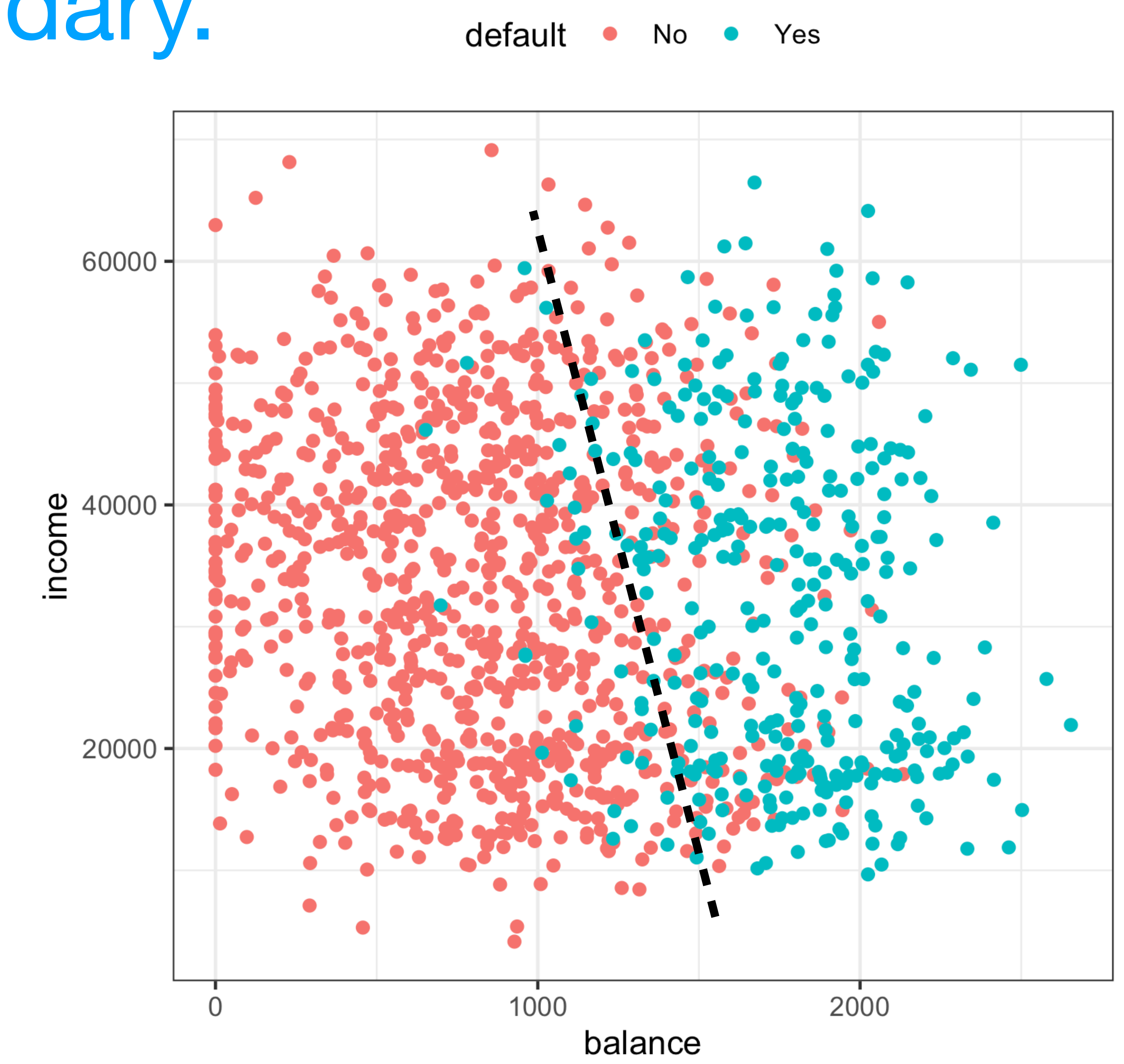
Classification via logistic regression

$$\text{default} = \begin{cases} \text{Yes}, & \text{if } \widehat{\mathbb{P}}[\text{default}] \geq c; \\ \text{No}, & \text{if } \widehat{\mathbb{P}}[\text{default}] < c. \end{cases}$$

$$\widehat{\mathbb{P}}[\text{default}] > 0.5 \iff \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{student} + \hat{\beta}_2 \cdot \text{balance} + \hat{\beta}_3 \cdot \text{income} > 0$$

Logistic regression has a linear decision boundary.





Separable data

When the two classes of response variable can be perfectly separated in feature space, logistic regression solution undefined, though perfect predictions possible.

A similar phenomenon occurs in linear regression under perfect multicollinearity: The coefficient estimates are undefined but good prediction still possible.