Linear and logistic regression

STAT 4710

October 4, 2022

Rolling into Unit 3

- **Unit 1:** R for data mining
- **Unit 2:** Prediction fundamentals
- **Unit 3:** Regression-based methods
- **Unit 4:** Tree-based methods
- **Unit 5:** Deep learning

- **Lecture 1:** Linear and logistic regression
- **Lecture 2:** Regression in high dimensions
- **Lecture 3:** Ridge regression
- **Lecture 4:** Lasso regression
- **Lecture 5:** Unit review and quiz in class

Predicting a response based on multiple features

Predicting a response based on multiple features

If we want to predict income, we should not only use age! We might want to consider other factors like education, job type, sex, marital status, race, etc.

Predicting a response based on multiple features

If we want to predict income, we should not only use age! We might want to consider other factors like education, job type, sex, marital status, race, etc.

Given features $X_1, X_2, \ldots, X_{p-1}$, the most common way to model a response Y is the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon.$$

Predicting a response based on multiple features

If we want to predict income, we should not only use age! We might want to consider other factors like education, job type, sex, marital status, race, etc.

Given features $X_1, X_2, \ldots, X_{p-1}$, the most common way to model a response Y is the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon.$$

Let's review:

- Continuous and categorical features in linear models
- Interpretation of linear regression coefficients
- How to fit a linear regression model

Continuous and categorical features in linear regression

Continuous and categorical features in linear regression

Features X_j must be expressed as numbers for $\beta_j X_j$ to make sense.

Continuous and categorical features in linear regression

Features X_j must be expressed as numbers for $\beta_j X_j$ to make sense.

Example 1 (continuous feature): $X_1 = \text{age}$. Continuous features are already numbers, so it makes sense to write $\beta_1 X_1$.

Continuous and categorical features in linear regression

Features X_j must be expressed as numbers for $\beta_j X_j$ to make sense.

Example 1 (continuous feature): $X_1 =$ age. Continuous features are already numbers, so it makes sense to write $\beta_1 X_1$.

Example 2 (binary feature): $X_2 =$ sex. It does not make sense to write $\beta_2 X_2$; what does $3 \times$ "male" mean? Instead, use dummy coding: $X_2 = I(\text{sex} = \text{male})$.

Continuous and categorical features in linear regression

Features X_j must be expressed as numbers for $\beta_j X_j$ to make sense.

Example 1 (continuous feature): $X_1 =$ age. Continuous features are already numbers, so it makes sense to write $\beta_1 X_1$.

Example 2 (binary feature): $X_2 =$ sex. It does not make sense to write $\beta_2 X_2$; what does $3 \times$ "male" mean? Instead, use dummy coding: $X_2 = I($sex $=$ male$)$.

Example 3 (categorical feature): $X_3 =$ education. It does not make sense to write $\beta_3 X_3$. Instead, map education onto multiple dummy variables: $X_3 = I($education = high school$)$, $X_4 = I($education = "college"$)$, etc.

Continuous and categorical features in linear regression

Features X_j must be expressed as numbers for $\beta_j X_j$ to make sense.

Example 1 (continuous feature): $X_1 =$ age. Continuous features are already numbers, so it makes sense to write $\beta_1 X_1$.

Example 2 (binary feature): $X_2 =$ sex. It does not make sense to write $\beta_2 X_2$; what does $3 \times$ "male" mean? Instead, use dummy coding: $X_2 = I(\text{sex} = \text{male})$.

Example 3 (categorical feature): $X_3 =$ education. It does not make sense to write $\beta_3 X_3$. Instead, map education onto multiple dummy variables:
$X_3 = I(\text{education = high school})$, $X_4 = I(\text{education = "college"})$, etc.

To avoid redundancy, use dummy variables for all levels except one baseline.

Interpretation of linear regression coefficients

Interpretation of linear regression coefficients

Consider the following linear regression model:

Interpretation of linear regression coefficients

Consider the following linear regression model:

$$\text{income} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot I(\text{sex = "M"}) + \beta_3 \cdot I(\text{ed = "HS"}) + \beta_4 \cdot I(\text{ed = "college"}) + \epsilon$$

Interpretation of linear regression coefficients

Consider the following linear regression model:

$$\text{income} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot I(\text{sex = "M"}) + \beta_3 \cdot I(\text{ed = "HS"}) + \beta_4 \cdot I(\text{ed = "college"}) + \epsilon$$

Example 1 (continuous feature): β_1 represents increase in mean income associated with extra year of age.

Interpretation of linear regression coefficients

Consider the following linear regression model:

$$\text{income} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot I(\text{sex = "M"}) + \beta_3 \cdot I(\text{ed = "HS"}) + \beta_4 \cdot I(\text{ed = "college"}) + \epsilon$$

Example 1 (continuous feature): β_1 represents increase in mean income associated with extra year of age.

Example 2 (binary feature): β_2 represents increase in mean income associated with moving from female (baseline) to male.

Interpretation of linear regression coefficients

Consider the following linear regression model:

$$\text{income} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot I(\text{sex = "M"}) + \beta_3 \cdot I(\text{ed = "HS"}) + \beta_4 \cdot I(\text{ed = "college"}) + \epsilon$$

Example 1 (continuous feature): β_1 represents increase in mean income associated with extra year of age.

Example 2 (binary feature): β_2 represents increase in mean income associated with moving from female (baseline) to male.

Example 3 (categorical feature): β_3 represents increase in mean income associated with moving from less than HS education (baseline) to HS education.

Interpretation of linear regression coefficients

Consider the following linear regression model:

$$\text{income} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot I(\text{sex = "M"}) + \beta_3 \cdot I(\text{ed = "HS"}) + \beta_4 \cdot I(\text{ed = "college"}) + \epsilon$$

Example 1 (continuous feature): β_1 represents increase in mean income associated with extra year of age.

Example 2 (binary feature): β_2 represents increase in mean income associated with moving from female (baseline) to male.

Example 3 (categorical feature): β_3 represents increase in mean income associated with moving from less than HS education (baseline) to HS education.

Note: Linear regression coefficients do not necessarily imply causation.

Fitting linear regression via least squares

Fitting linear regression via least squares

We have training data points (X_i, Y_i) for $i = 1, \ldots, n$.

Fitting linear regression via least squares

We have training data points (X_i, Y_i) for $i = 1, \ldots, n$.

Given coefficients β, define prediction $f_\beta(X_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$.

Fitting linear regression via least squares

We have training data points (X_i, Y_i) for $i = 1, \dots, n$.

Given coefficients β, define prediction $f_\beta(X_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$.

Based on the training data, we want to find $\widehat{\beta}$ such that $Y_i \approx f_{\widehat{\beta}}(X_i)$:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2.$$

Fitting linear regression via least squares

We have training data points (X_i, Y_i) for $i = 1, \ldots, n$.

Given coefficients β, define prediction $f_\beta(X_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$.

Based on the training data, we want to find $\widehat{\beta}$ such that $Y_i \approx f_{\widehat{\beta}}(X_i)$:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2.$$

This is the method of least squares, or ordinary least squares (OLS).

Fitting linear regression via least squares

We have training data points (X_i, Y_i) for $i = 1,\ldots,n$.

Given coefficients β, define prediction $f_\beta(X_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$.

Based on the training data, we want to find $\widehat{\beta}$ such that $Y_i \approx f_{\widehat{\beta}}(X_i)$:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(X_i))^2.$$

This is the method of least squares, or ordinary least squares (OLS).

The least squares optimization problem can be solved in closed form.

What if the response is binary?

```
> 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
```

What if the response is binary?

```
> 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?

What if the response is binary?

```
> 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

$$\mathbb{P}[\text{default} = \text{Yes} \mid \text{student, balance, income}]$$

and then predict

$$\text{default} = \begin{cases} \text{Yes}, & \text{if } \widehat{\mathbb{P}}[\text{default}] \geq 0.5; \\ \text{No}, & \text{if } \widehat{\mathbb{P}}[\text{default}] < 0.5. \end{cases}$$

What if the response is binary?

```
> 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

$$\mathbb{P}[\text{default} = \text{Yes} \mid \text{student, balance, income}]$$

and then predict

$$\text{default} = \begin{cases} \text{Yes}, & \text{if } \widehat{\mathbb{P}}[\text{default}] \geq 0.5; \\ \text{No}, & \text{if } \widehat{\mathbb{P}}[\text{default}] < 0.5. \end{cases}$$

How do we model probability of default?

Options for modeling probability of default

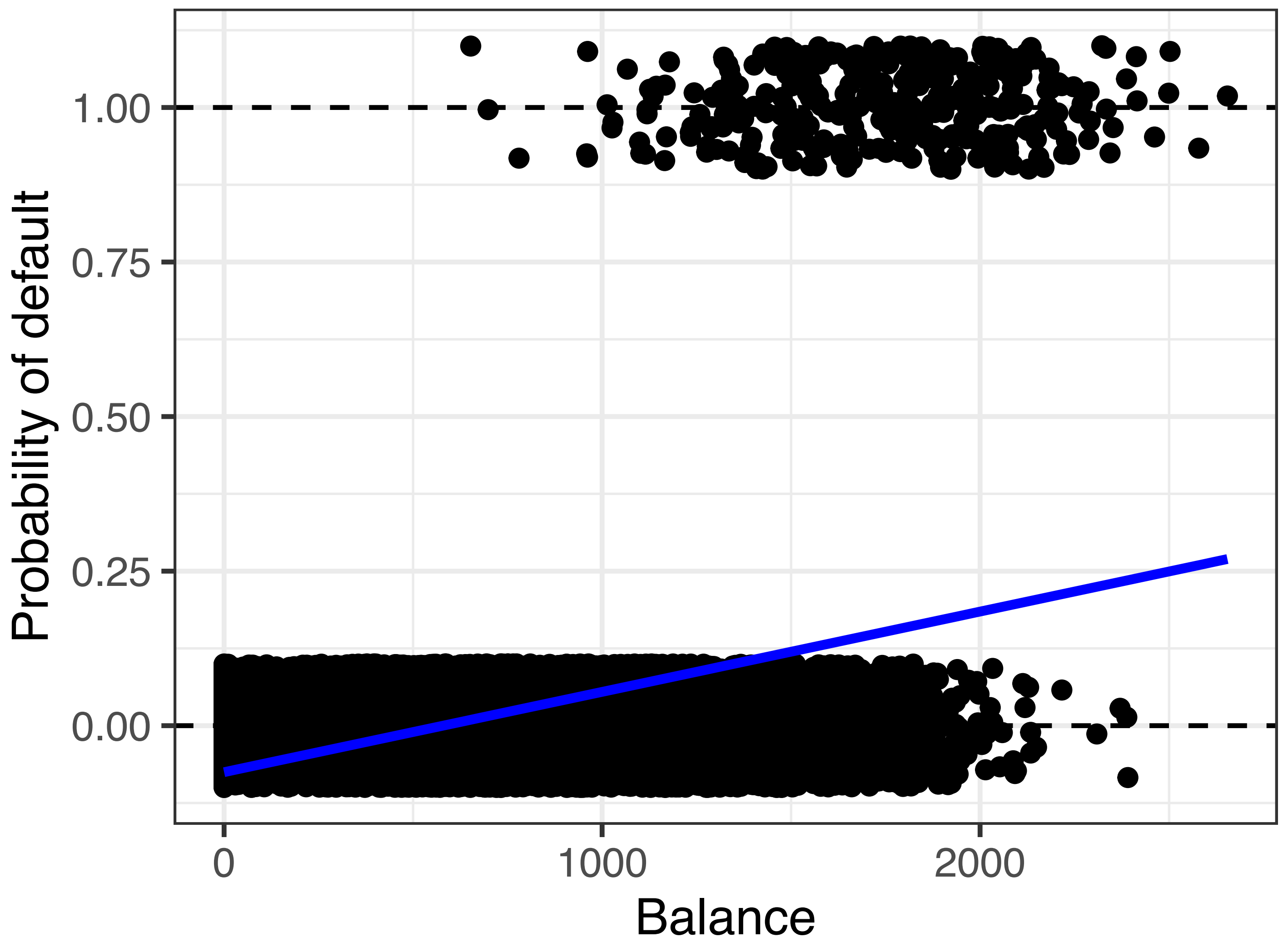
Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

Linear regression

$\beta_0 + \beta_1 \cdot \text{balance}$

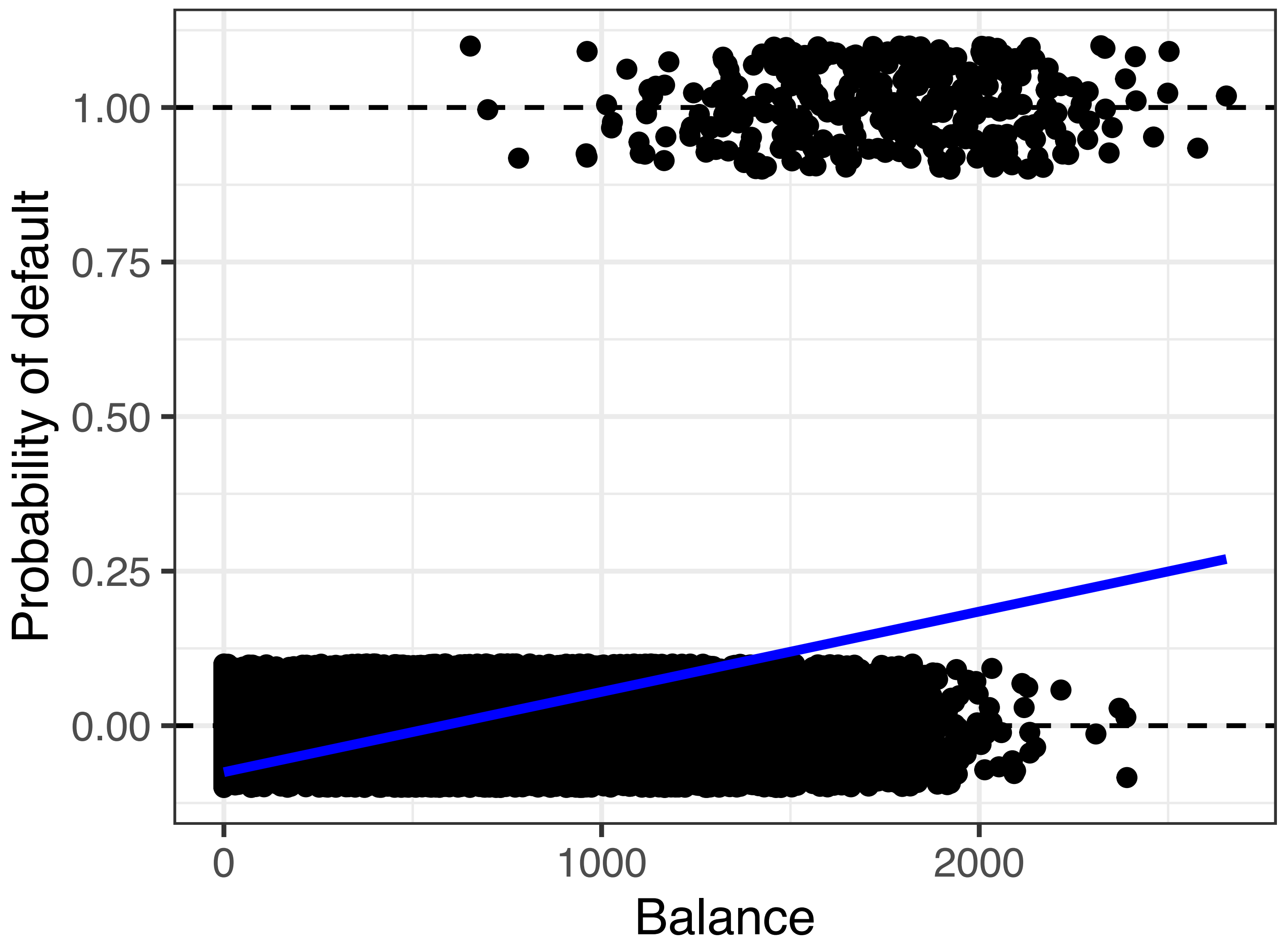


Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

Linear regression

$\beta_0 + \beta_1 \cdot \text{balance}$



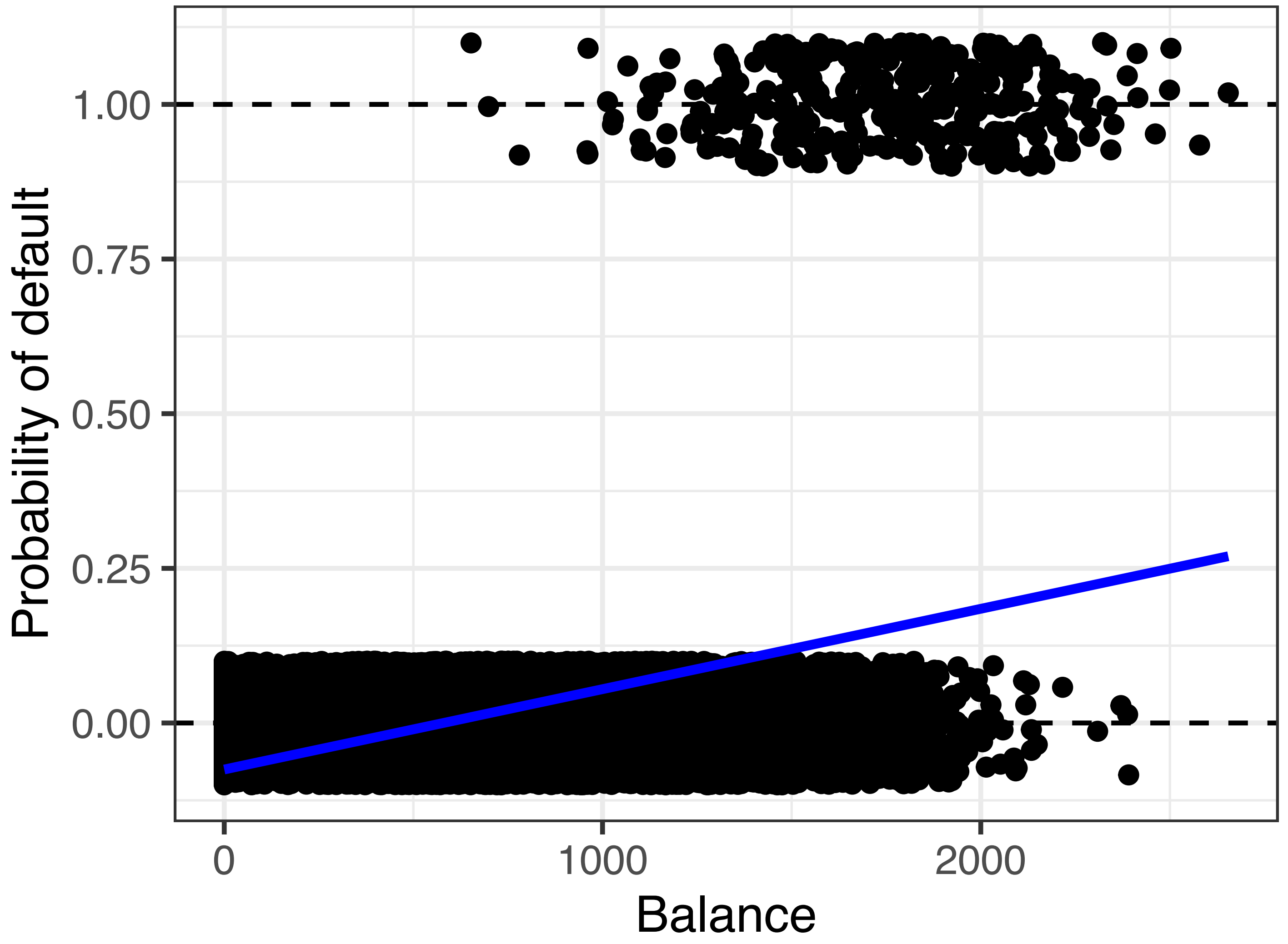
✔ Interpretable coefficients

Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

Linear regression

$\beta_0 + \beta_1 \cdot \text{balance}$



✔ Interpretable coefficients

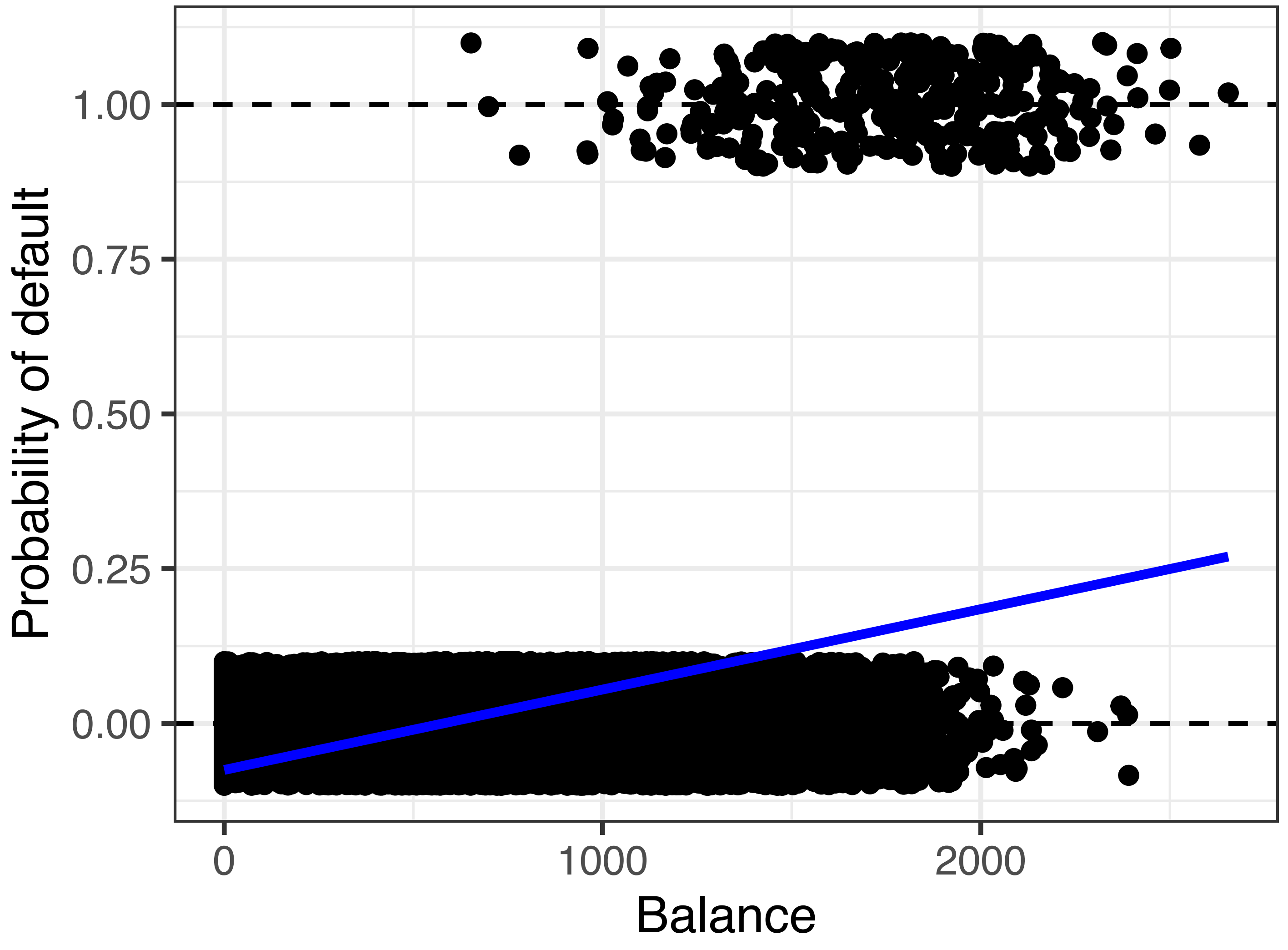
✘ Probabilities can fall outside [0,1]

Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

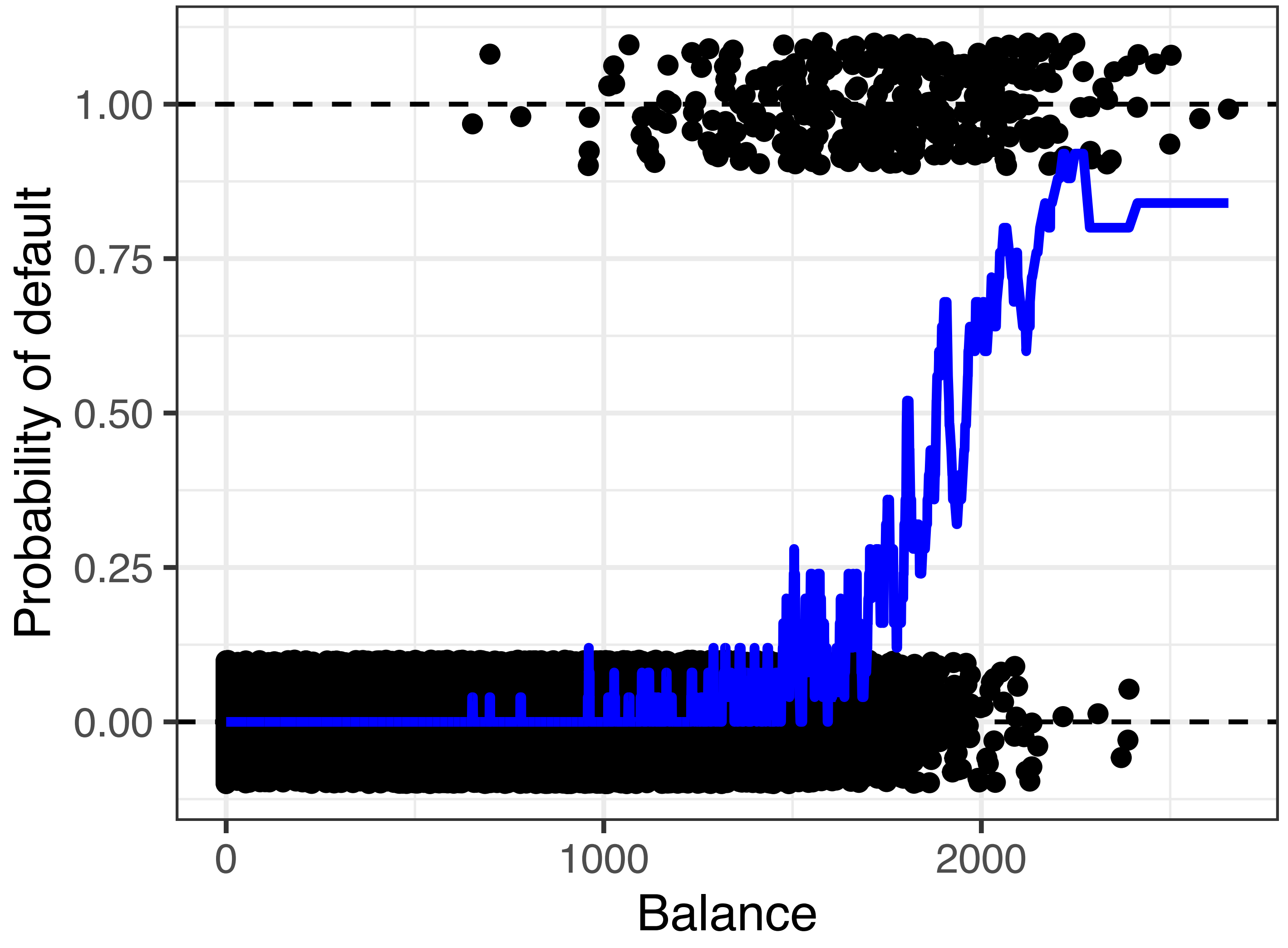
Linear regression

$\beta_0 + \beta_1 \cdot \text{balance}$



K-nearest neighbors

proportion of K N. N. who defaulted



✔ Interpretable coefficients

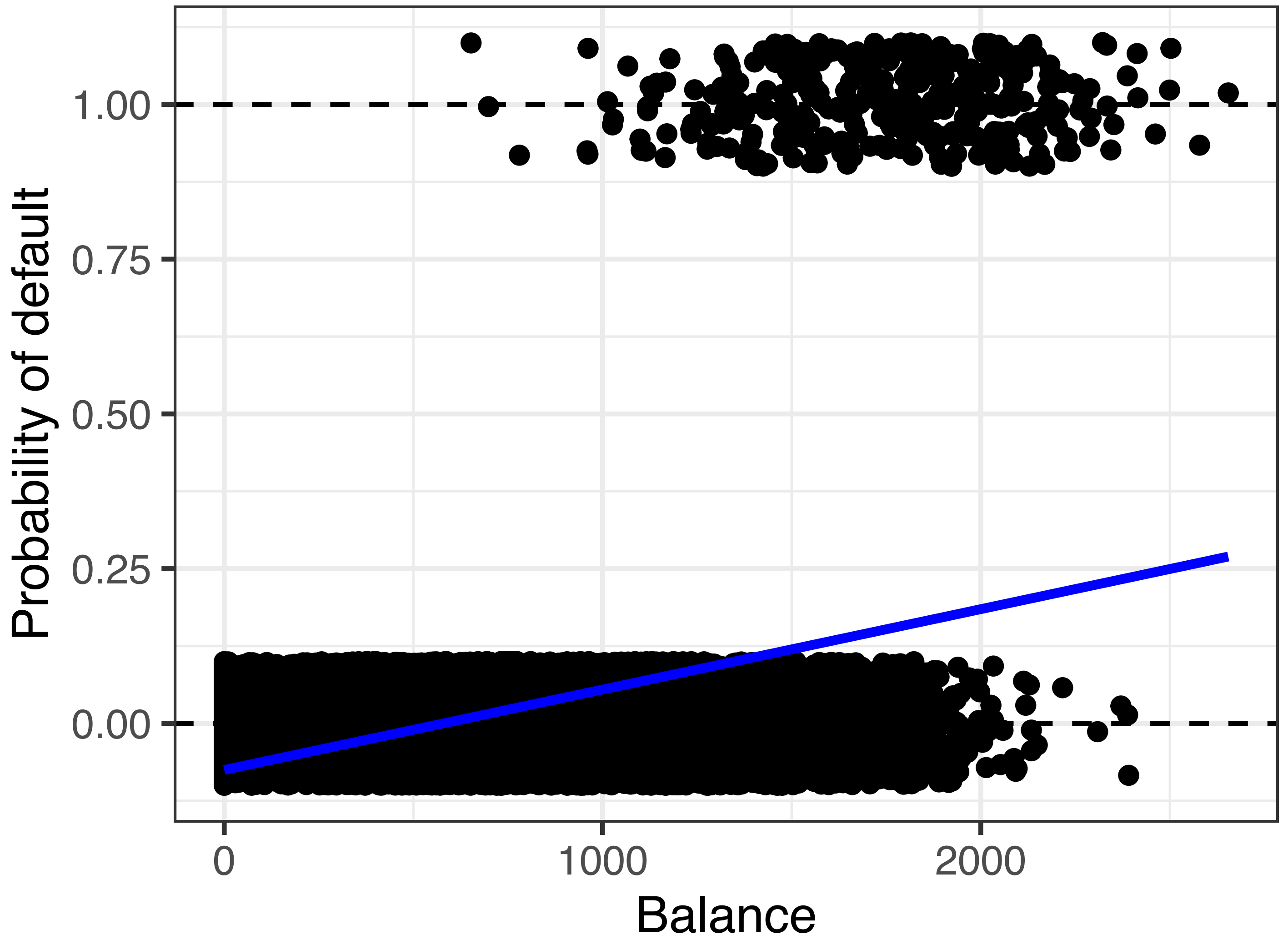
✘ Probabilities can fall outside [0,1]

Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

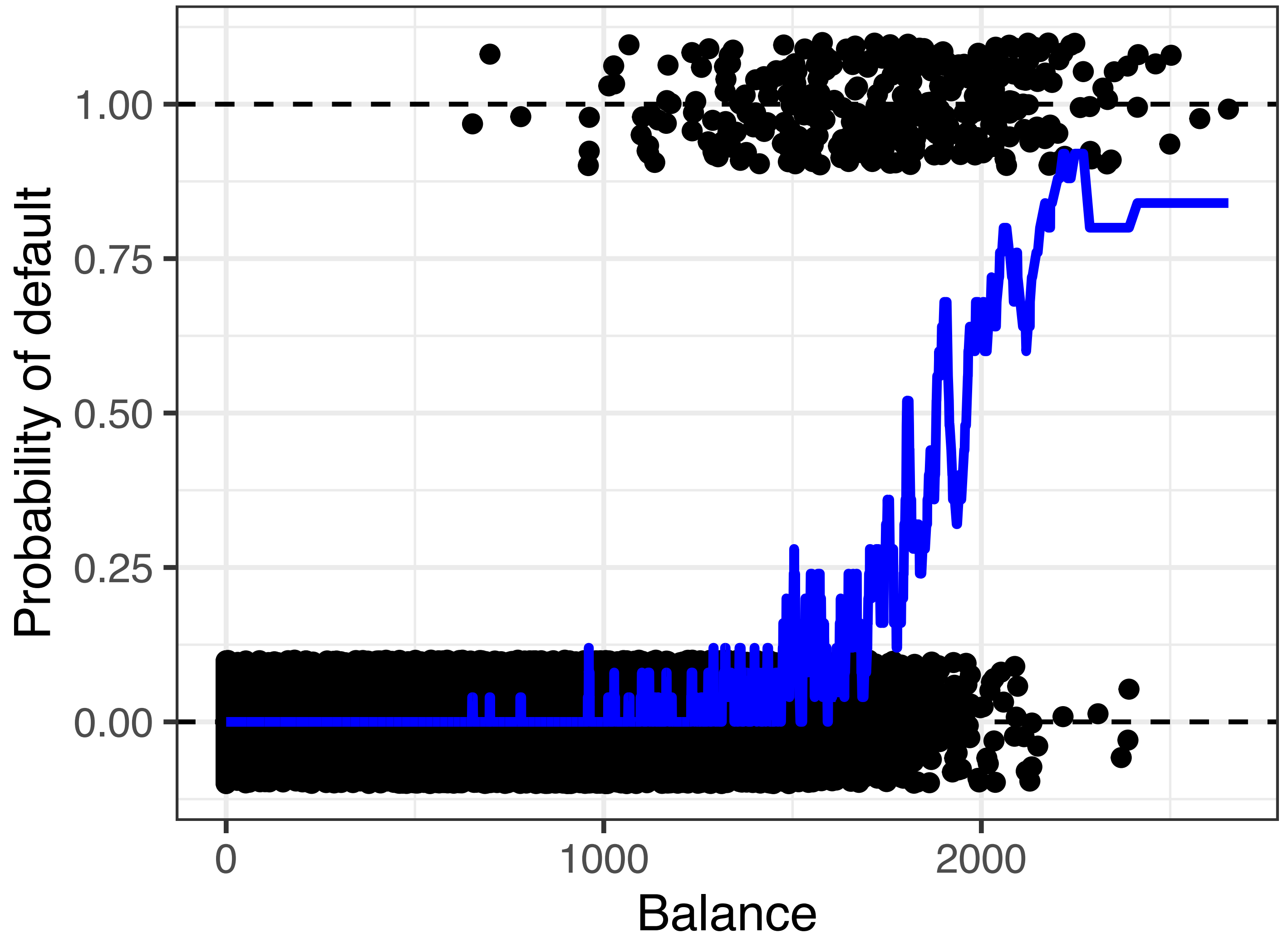
Linear regression

$\beta_0 + \beta_1 \cdot \text{balance}$



K-nearest neighbors

proportion of K N. N. who defaulted



✔ Interpretable coefficients

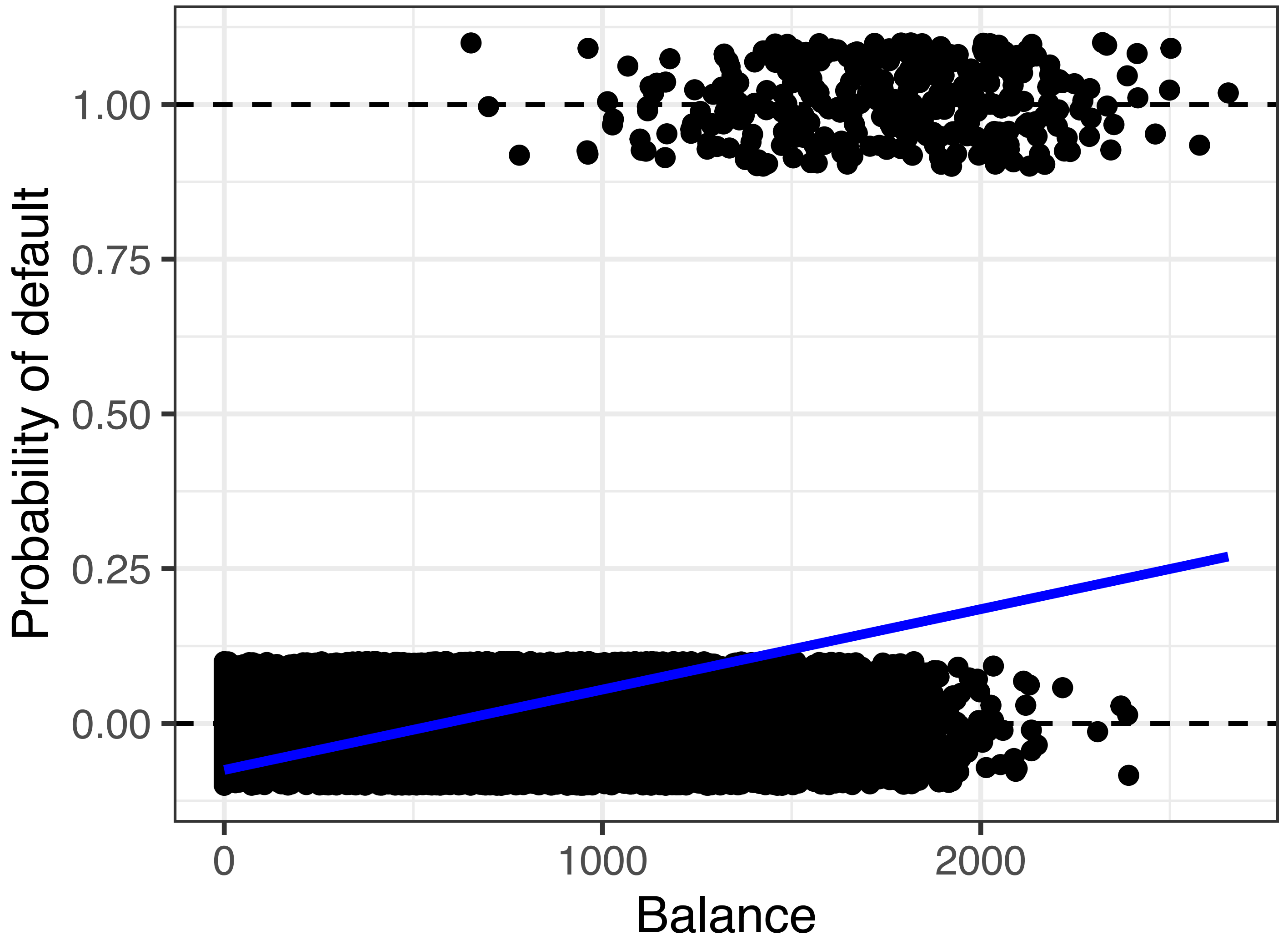
✘ Probabilities can fall outside [0,1]

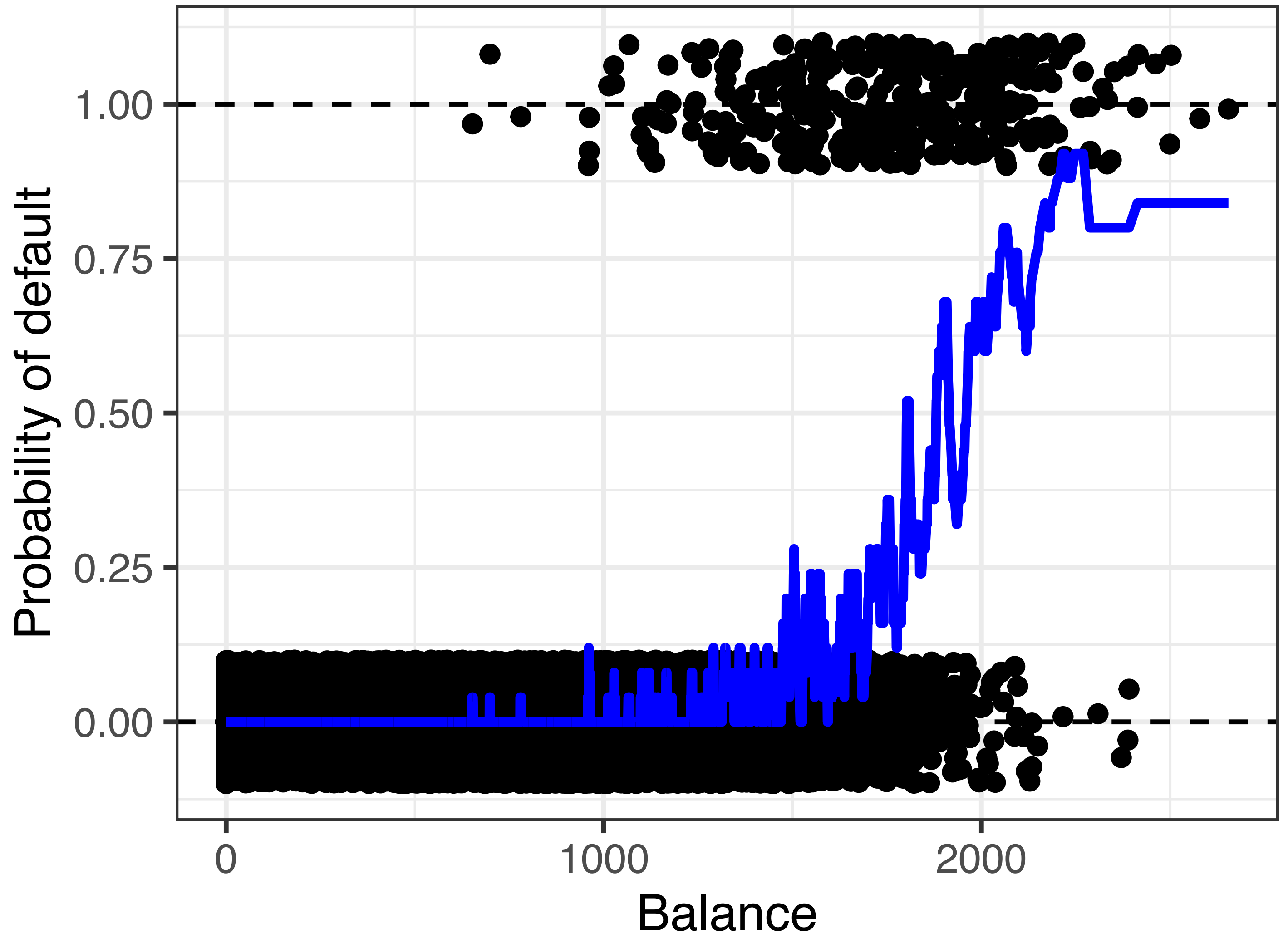
✘ Less interpretable model

Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

Linear regression	K-nearest neighbors
$\beta_0 + \beta_1 \cdot \text{balance}$	proportion of K N. N. who defaulted





✓ Interpretable coefficients

✗ Probabilities can fall outside [0,1]

✗ Less interpretable model

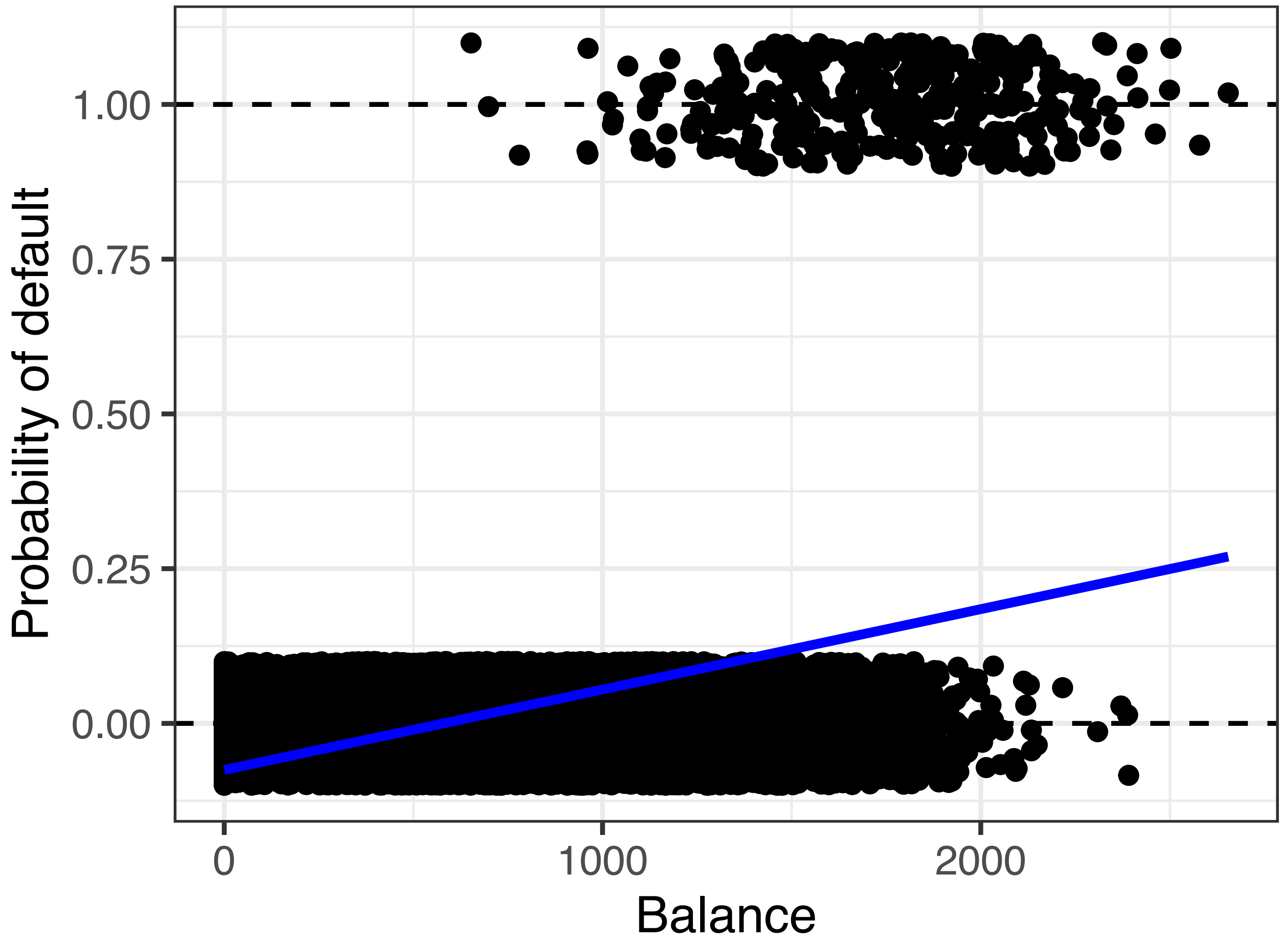
✓ Probabilities fall within [0,1]

Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

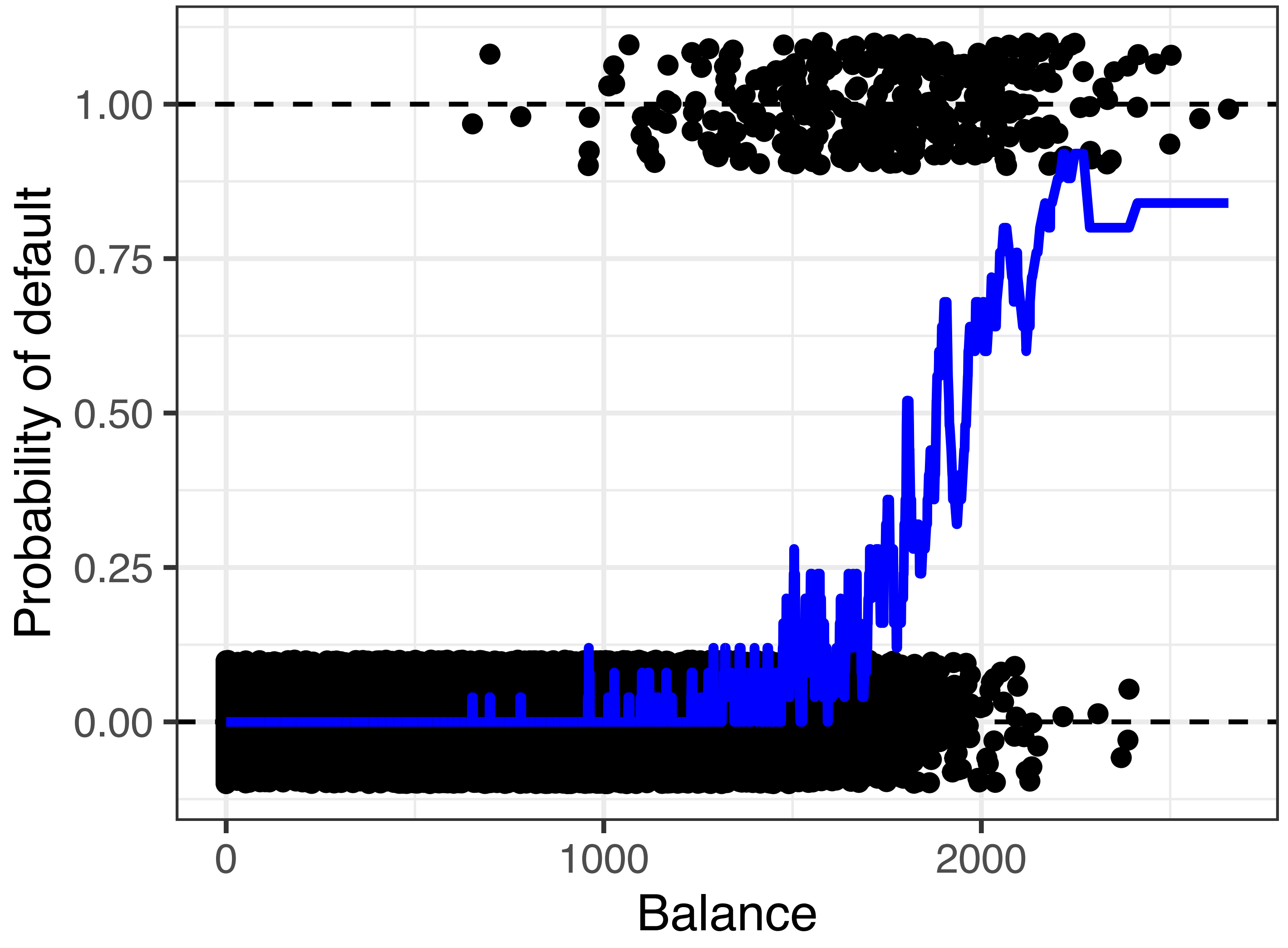
Linear regression

$\beta_0 + \beta_1 \cdot \text{balance}$



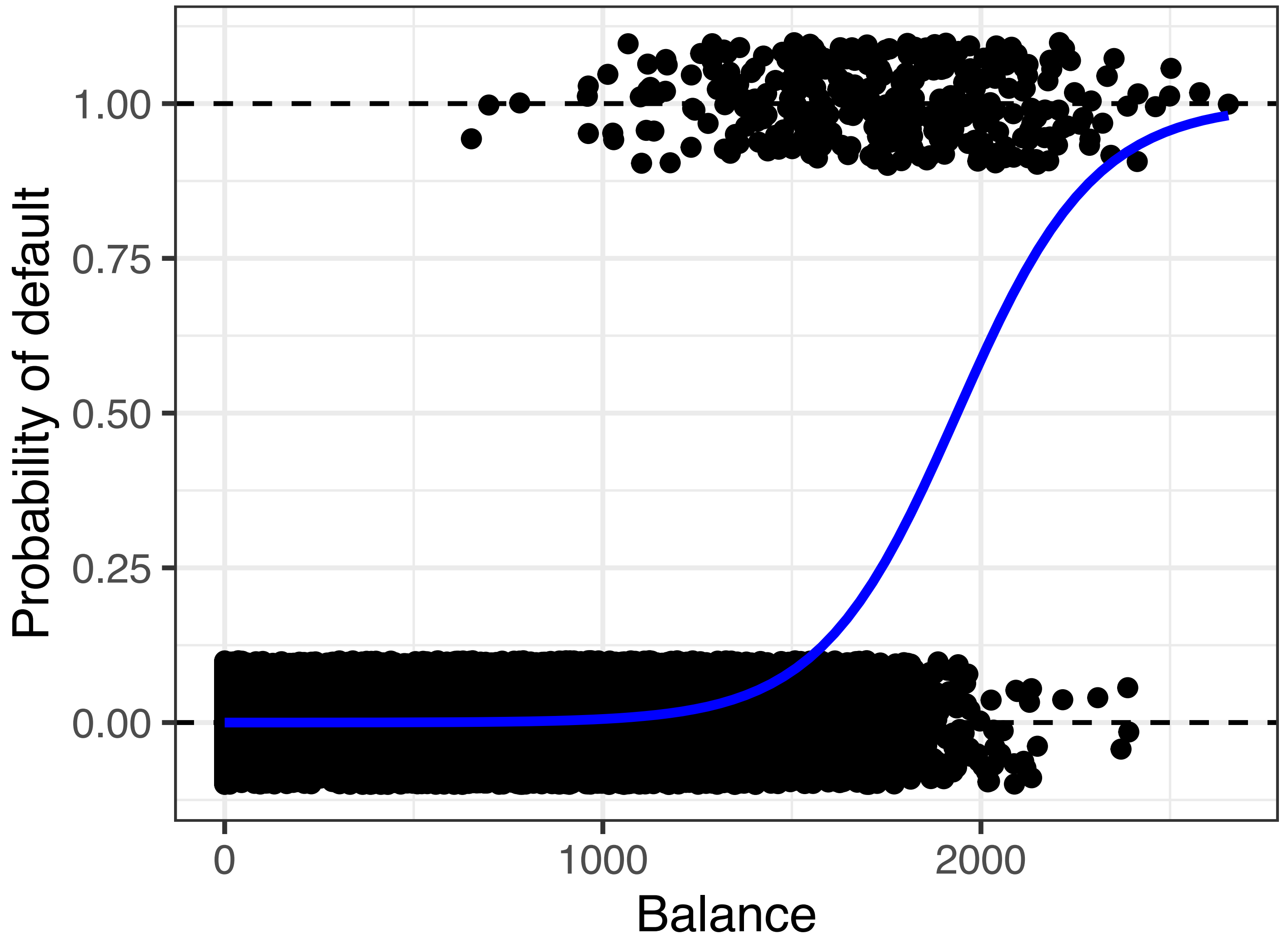
K-nearest neighbors

proportion of K N. N. who defaulted



Logistic regression

$\text{logistic}(\beta_0 + \beta_1 \cdot \text{balance})$



✓ Interpretable coefficients

✗ Probabilities can fall outside [0,1]

✗ Less interpretable model

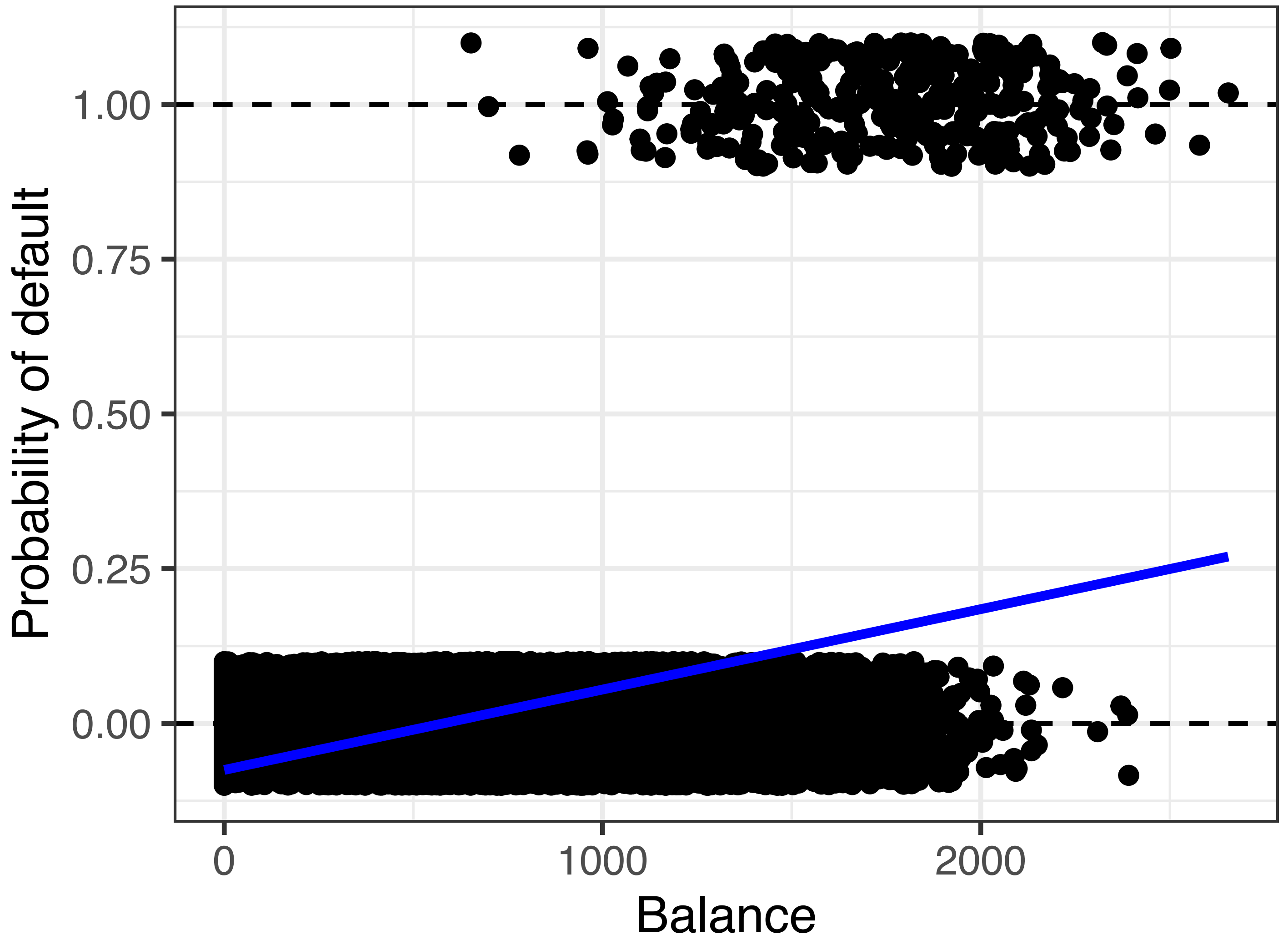
✓ Probabilities fall within [0,1]

Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

Linear regression

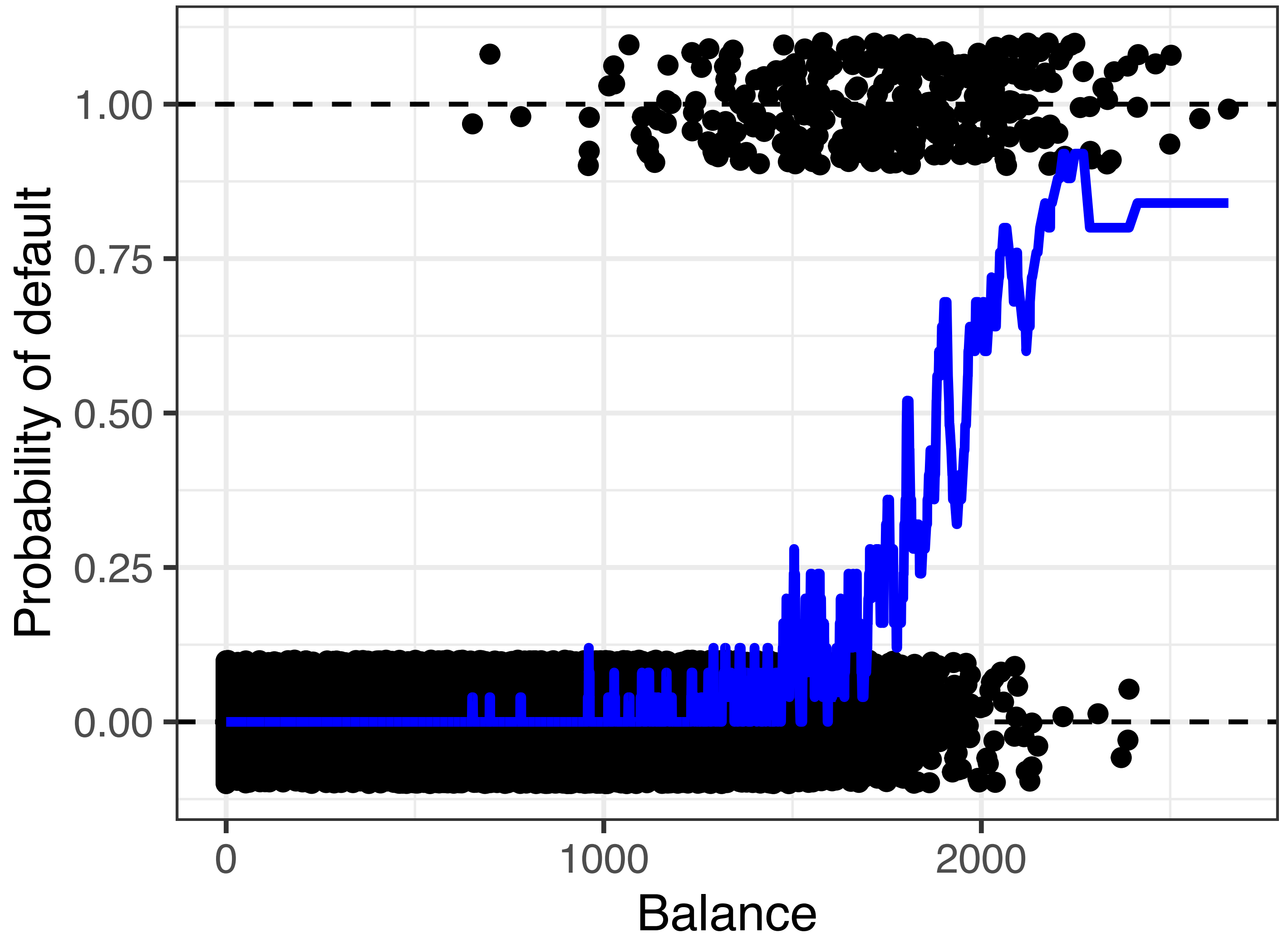
$\beta_0 + \beta_1 \cdot \text{balance}$



- ✓ Interpretable coefficients
- ✗ Probabilities can fall outside [0,1]

K-nearest neighbors

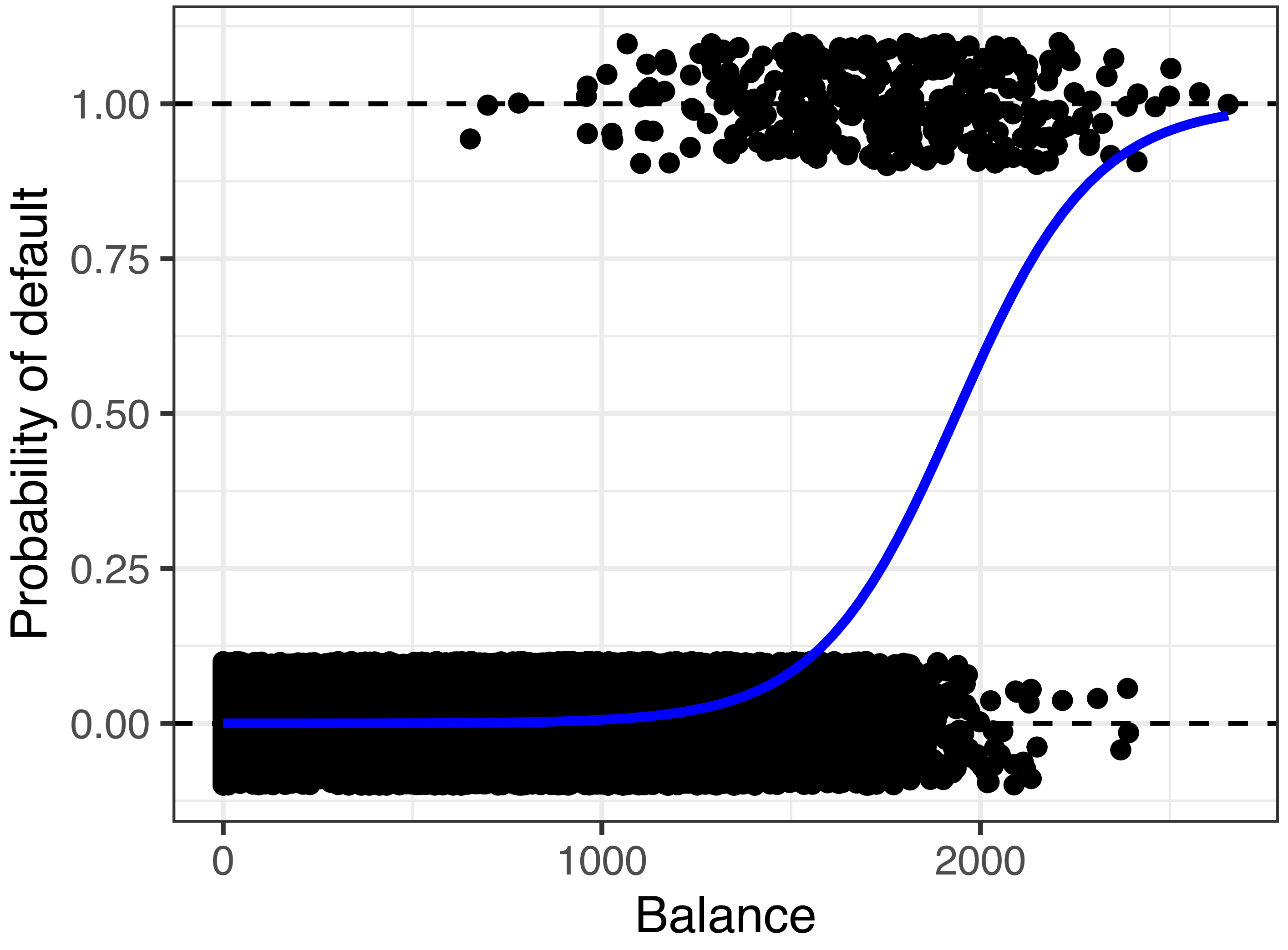
proportion of K N. N. who defaulted



- ✗ Less interpretable model
- ✓ Probabilities fall within [0,1]

Logistic regression

$\text{logistic}(\beta_0 + \beta_1 \cdot \text{balance})$

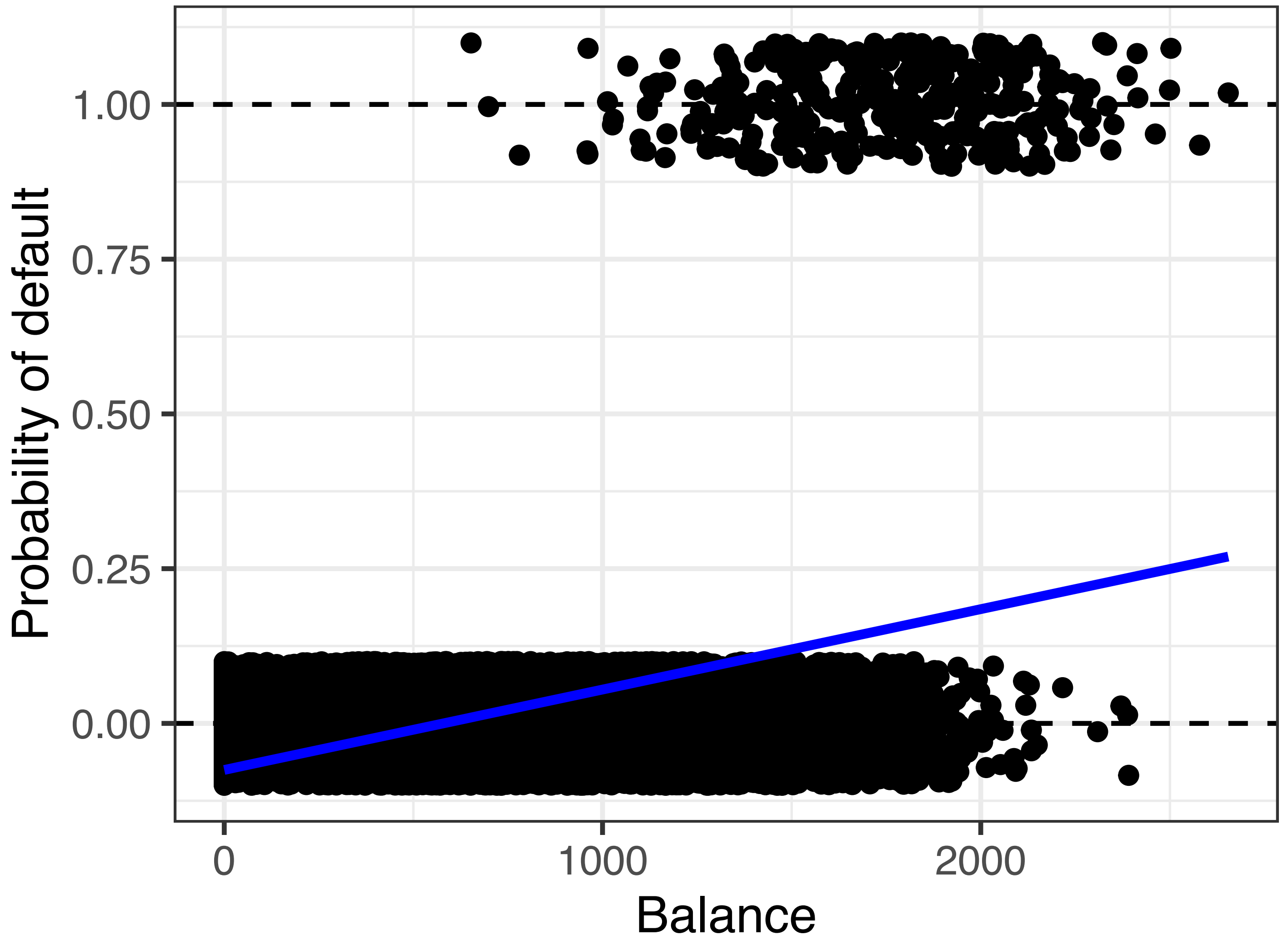


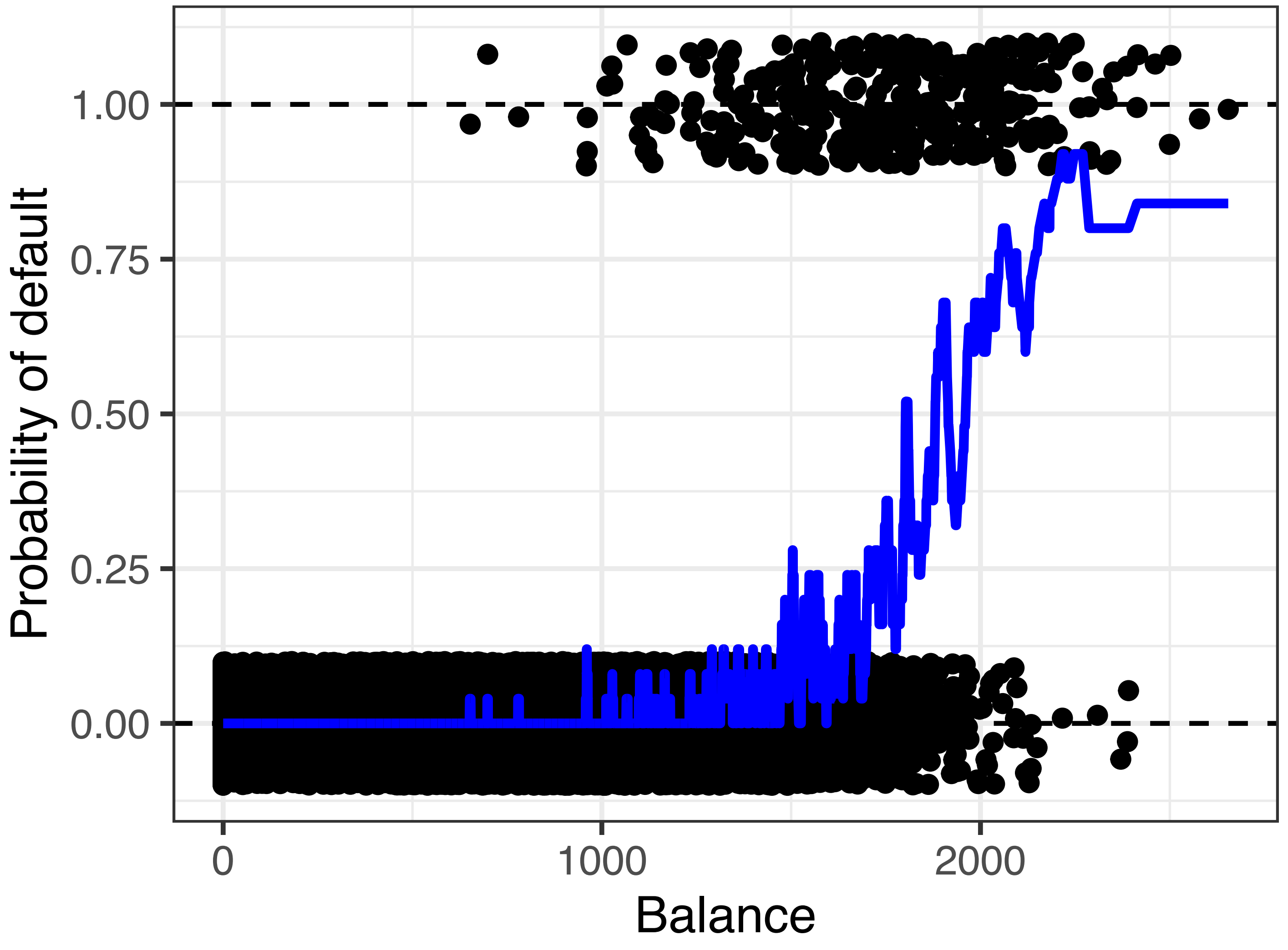
- ✓ Interpretable coefficients

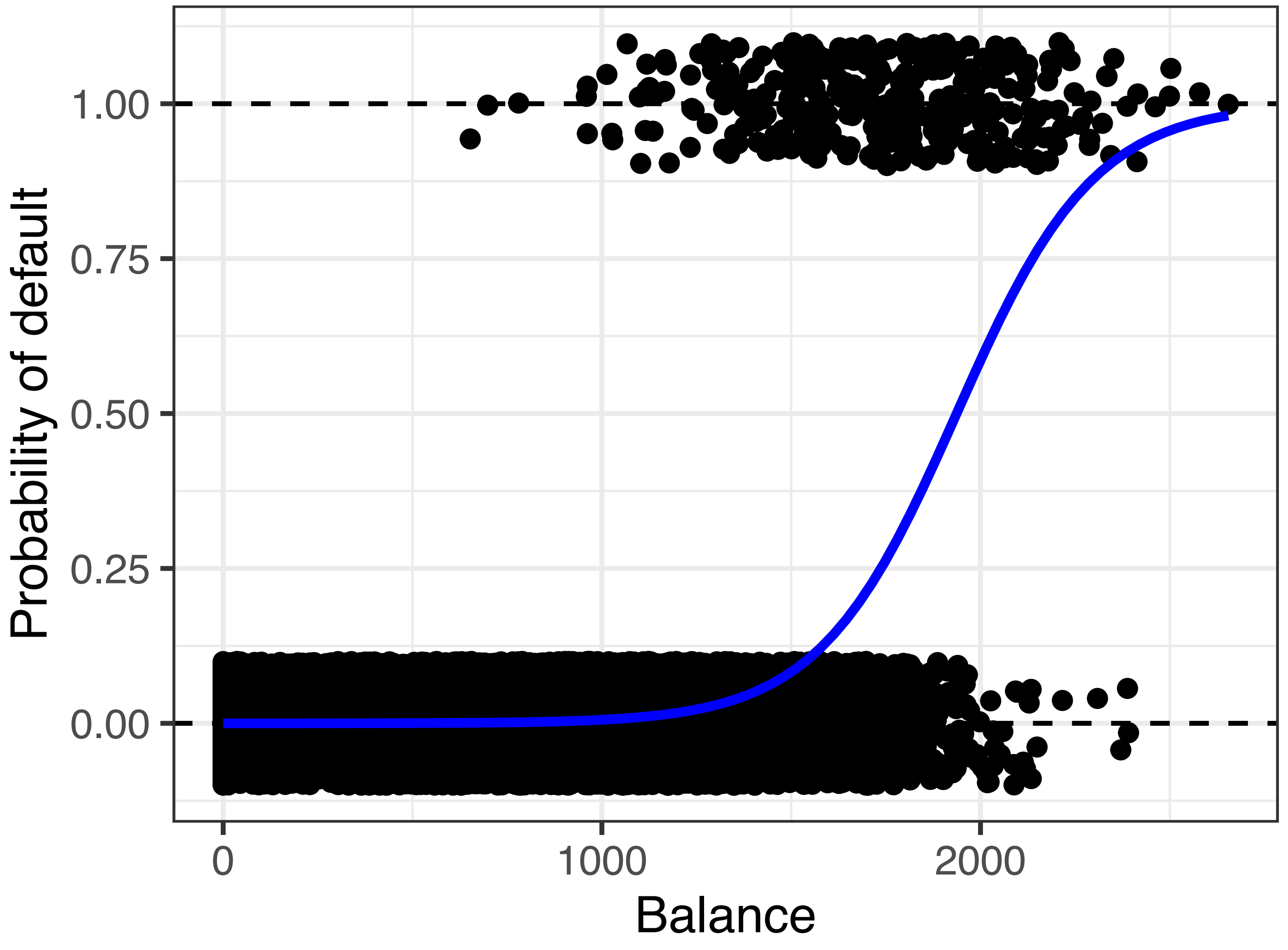
Options for modeling probability of default

Start by considering models for $\mathbb{P}[\text{default} \mid \text{balance}]$:

Linear regression	K-nearest neighbors	Logistic regression
$\beta_0 + \beta_1 \cdot \text{balance}$	proportion of K N. N. who defaulted	$\text{logistic}(\beta_0 + \beta_1 \cdot \text{balance})$







Linear regression	K-nearest neighbors	Logistic regression
✓ Interpretable coefficients	✗ Less interpretable model	✓ Interpretable coefficients
✗ Probabilities can fall outside [0,1]	✓ Probabilities fall within [0,1]	✓ Probabilities fall within [0,1]

The logistic regression model

The logistic regression model

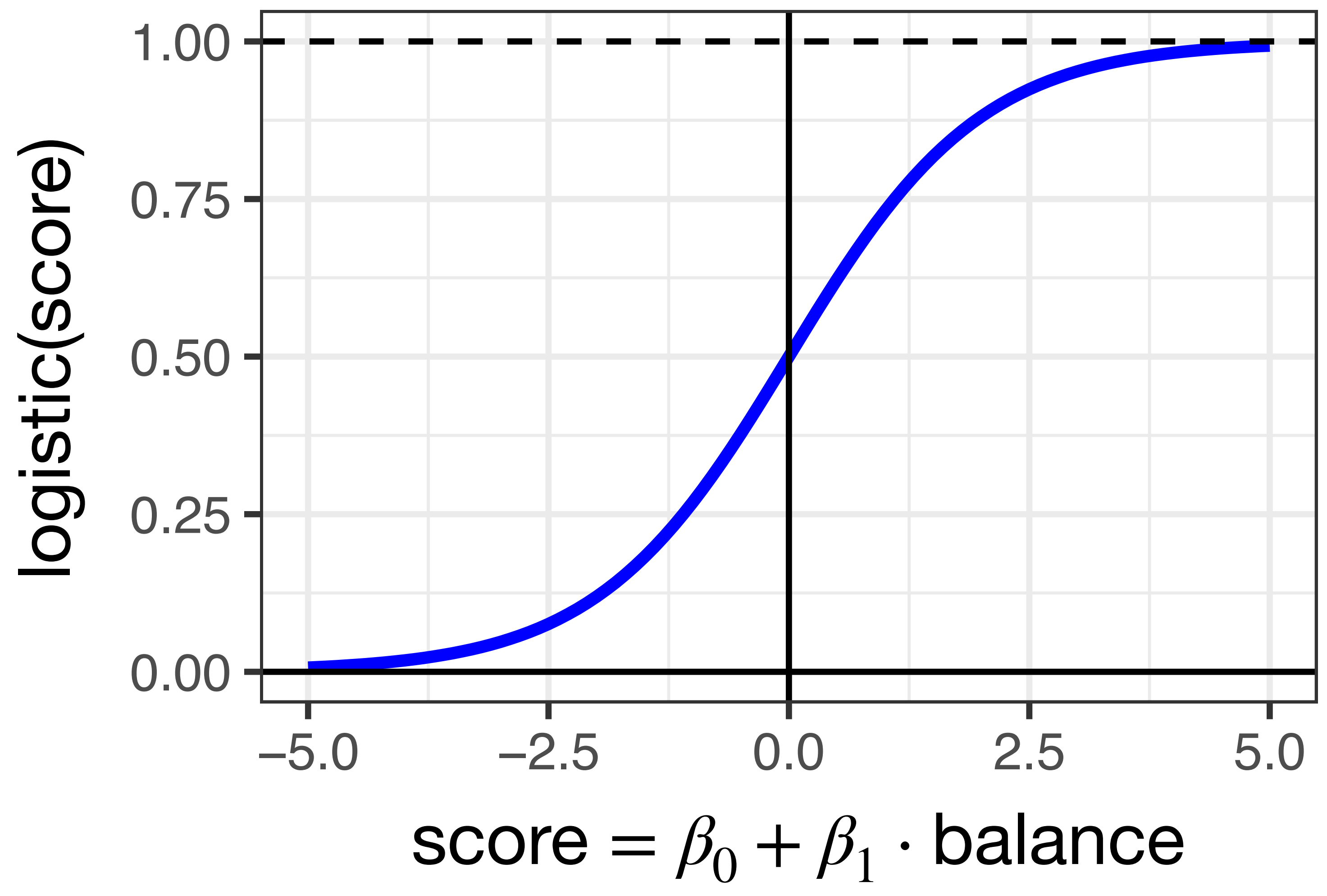
Use $\beta_0 + \beta_1 \cdot$ balance as a "score", then map the score onto [0,1] using logistic transformation:

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

The logistic regression model

Use $\beta_0 + \beta_1 \cdot$ balance as a "score", then map the score onto [0,1] using logistic transformation:

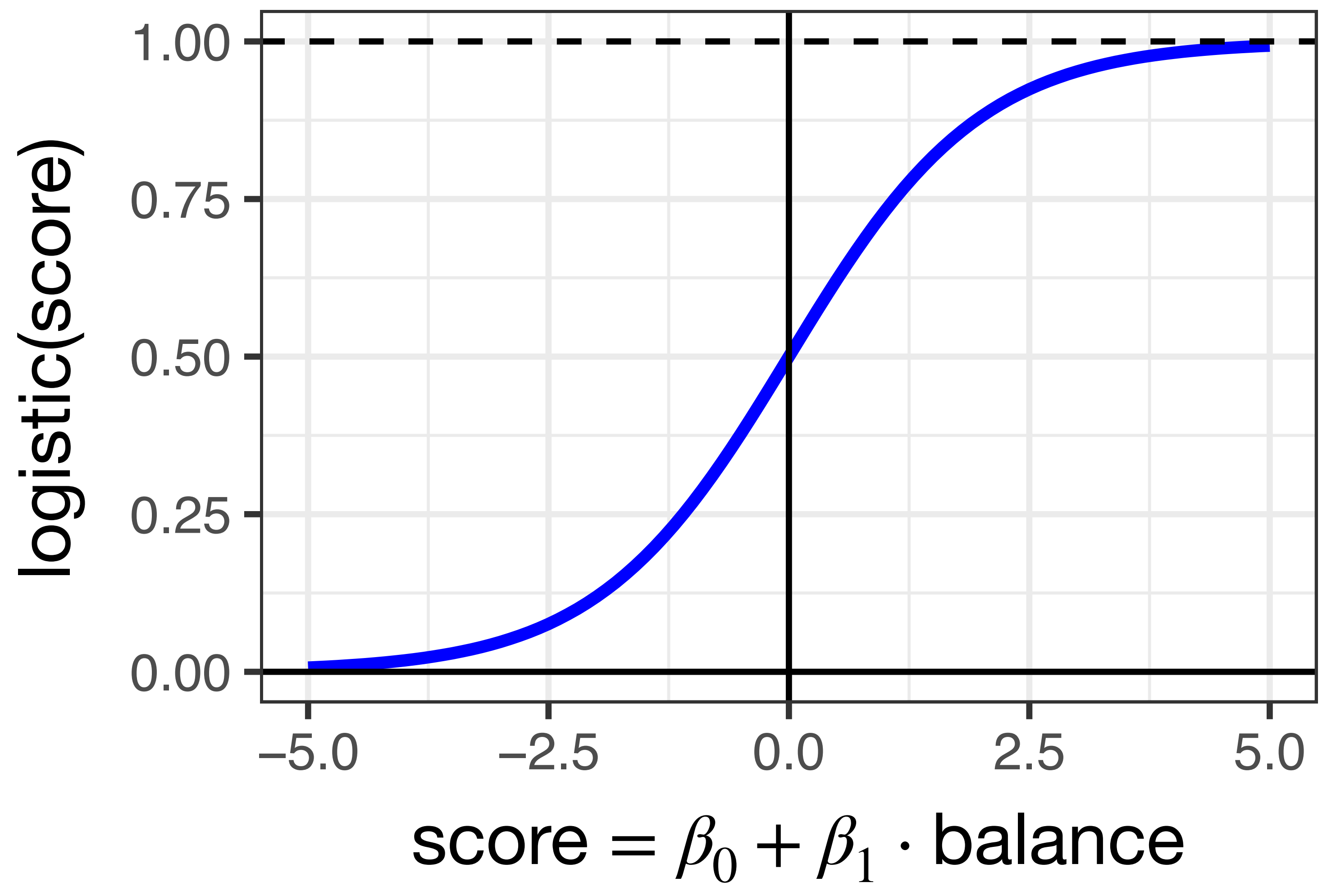
$$\text{logistic(score)} = \frac{e^{\text{score}}}{1 + e^{\text{score}}}$$



The logistic regression model

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

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



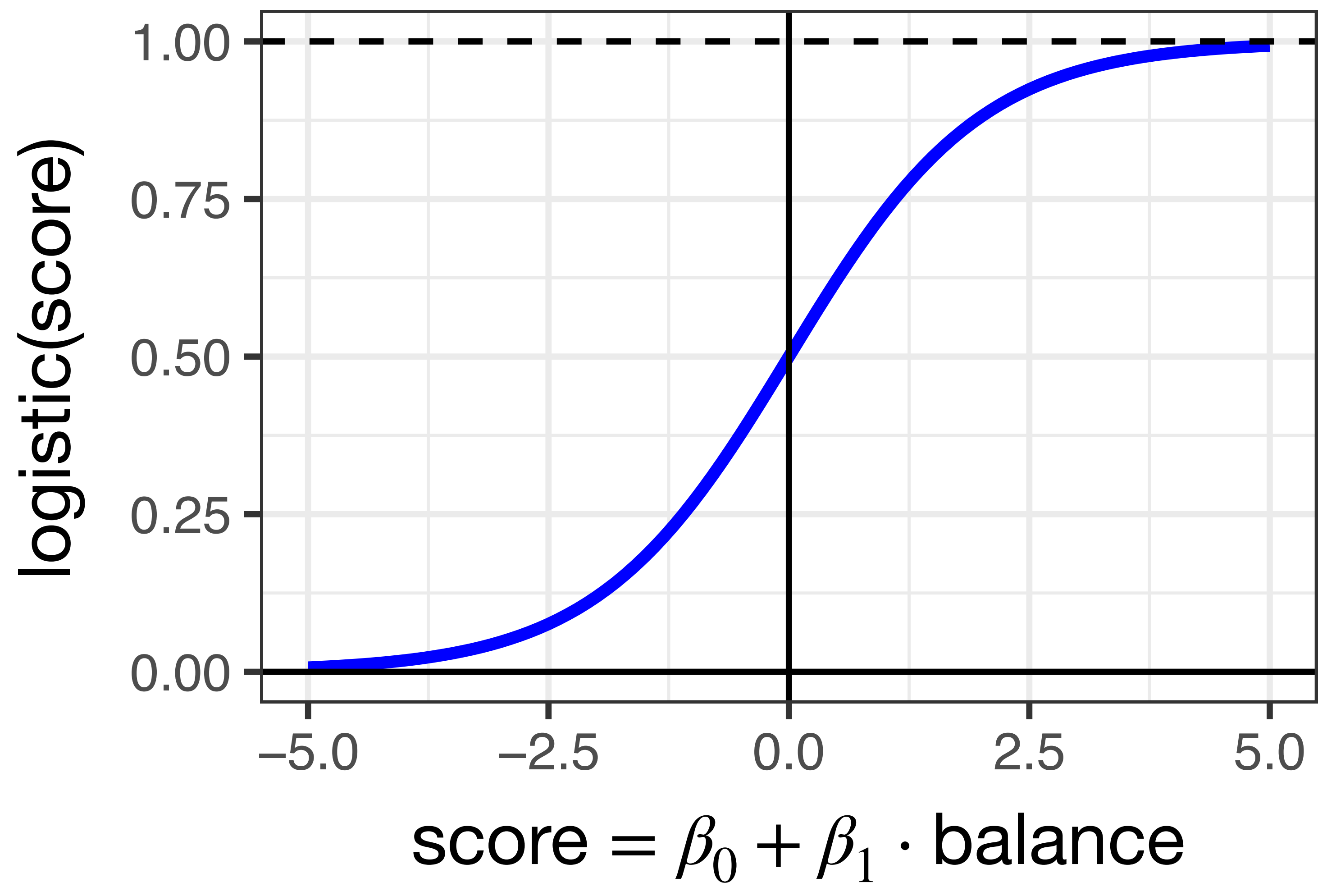
Logistic regression model:

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

The logistic regression model

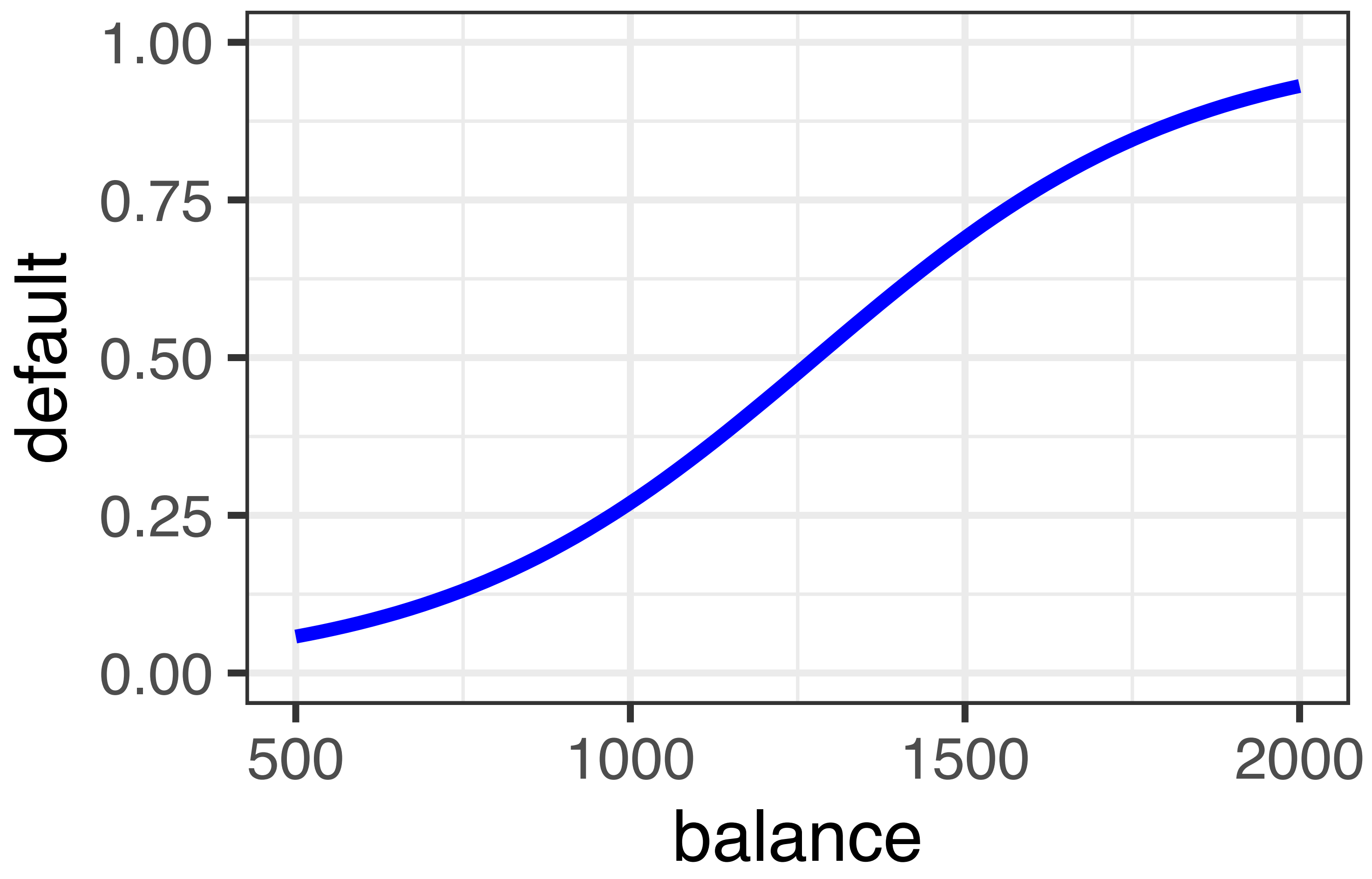
Use $\beta_0 + \beta_1 \cdot$ balance as a "score", then map the score onto [0,1] using logistic transformation:

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

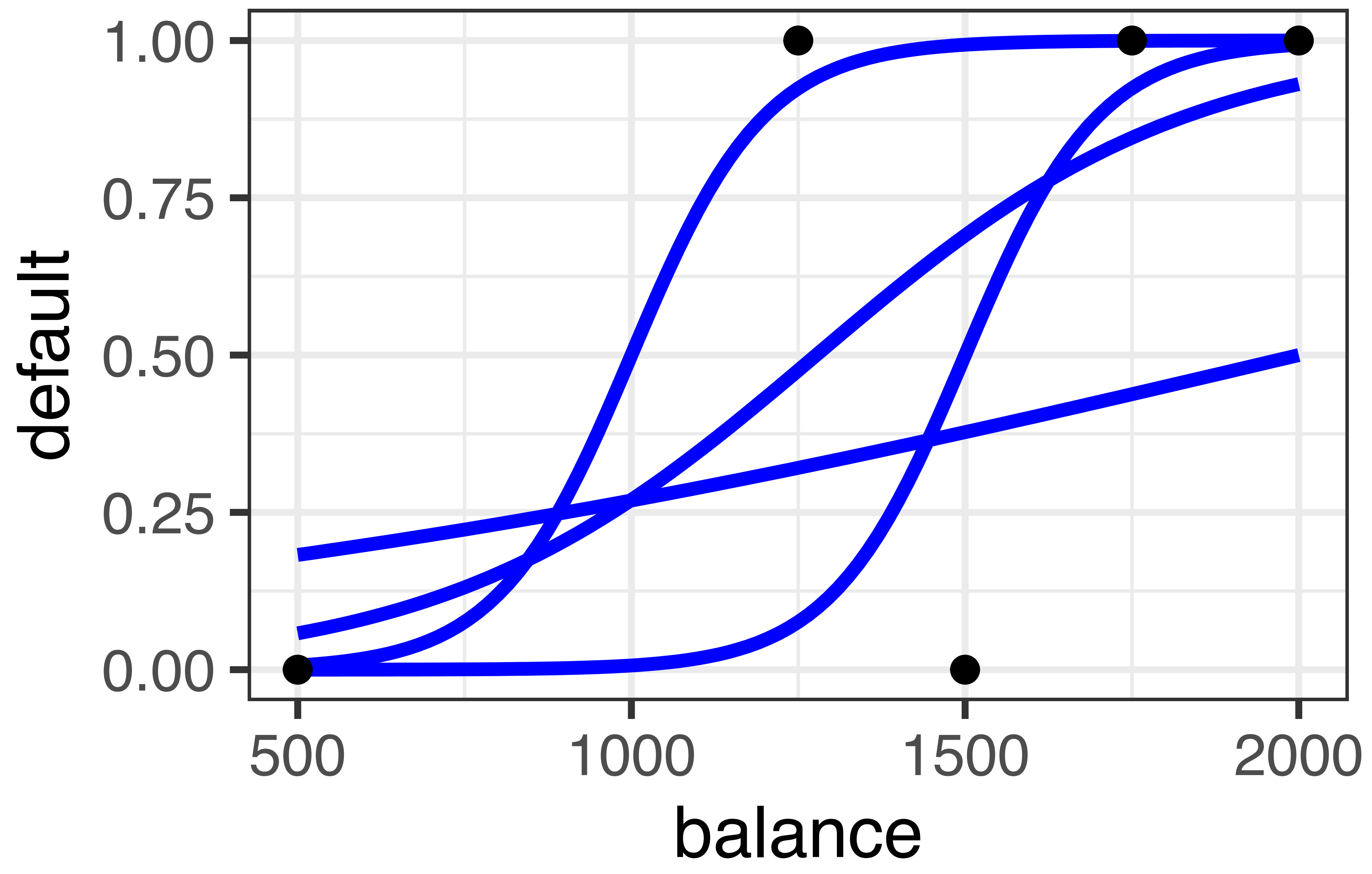


Logistic regression model:

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



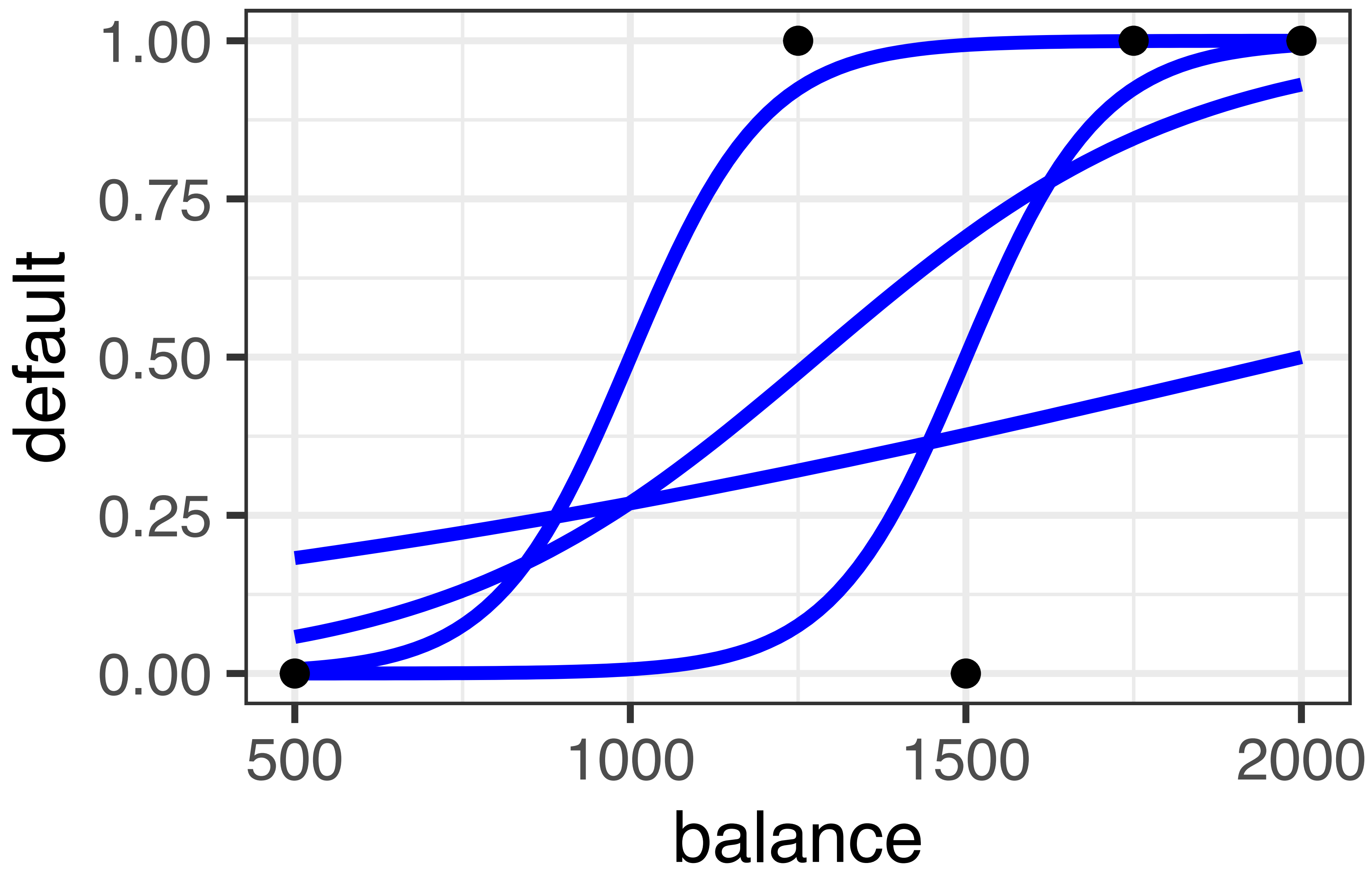
Fitting logistic regression models



Fitting logistic regression models

Each choice of (β_0, β_1) traces out a different logistic regression curve fit

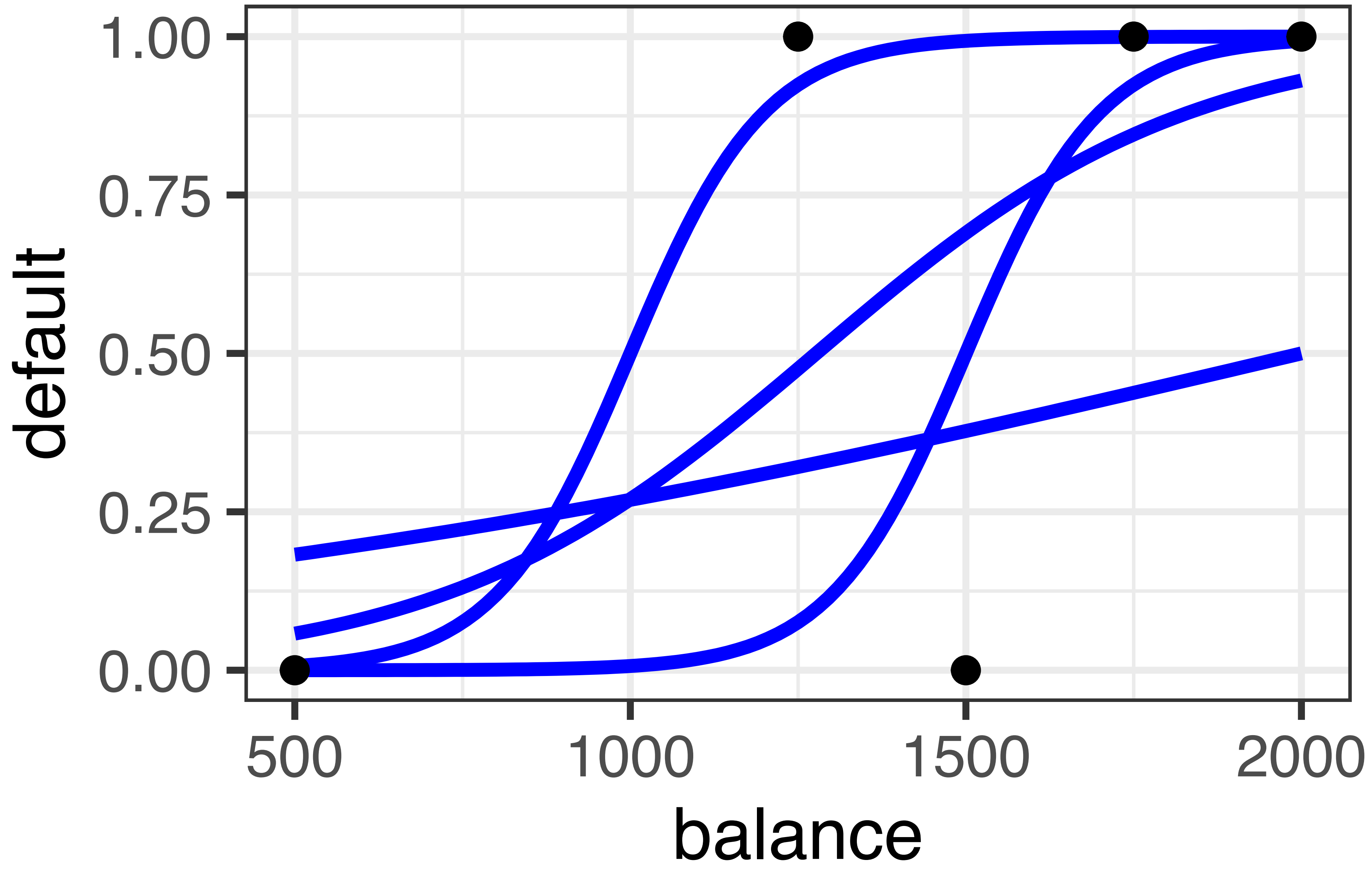
$$\mathbb{P}[\text{default} \mid \text{balance}] = \text{logistic}(\beta_0 + \beta_1 \cdot \text{balance}).$$



Fitting logistic regression models

Each choice of (β_0, β_1) traces out a different logistic regression curve fit

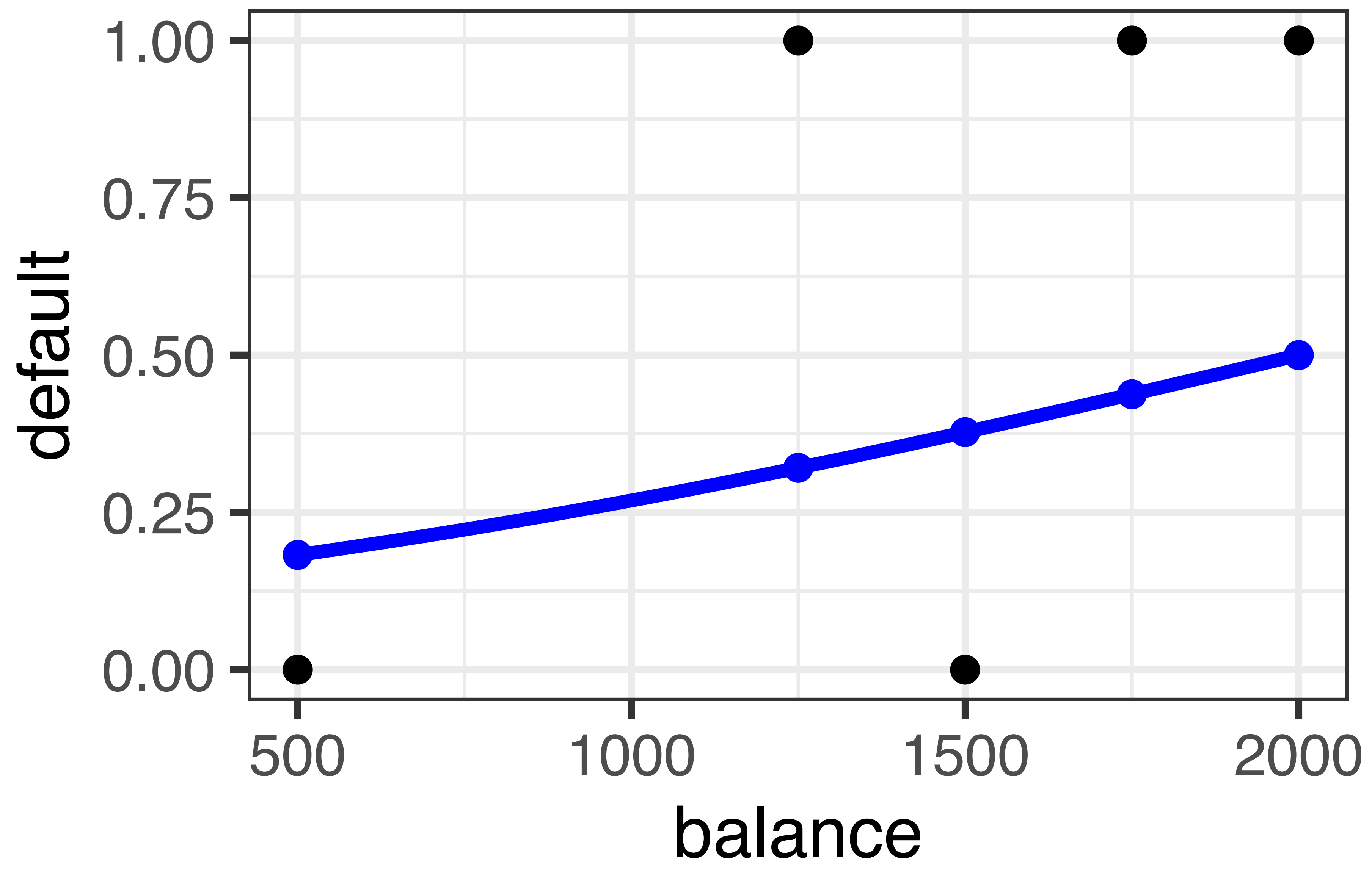
$$\mathbb{P}[\text{default} \mid \text{balance}] = \text{logistic}(\beta_0 + \beta_1 \cdot \text{balance}).$$



Which logistic regression curve fits the data the best?

Maximum likelihood estimation

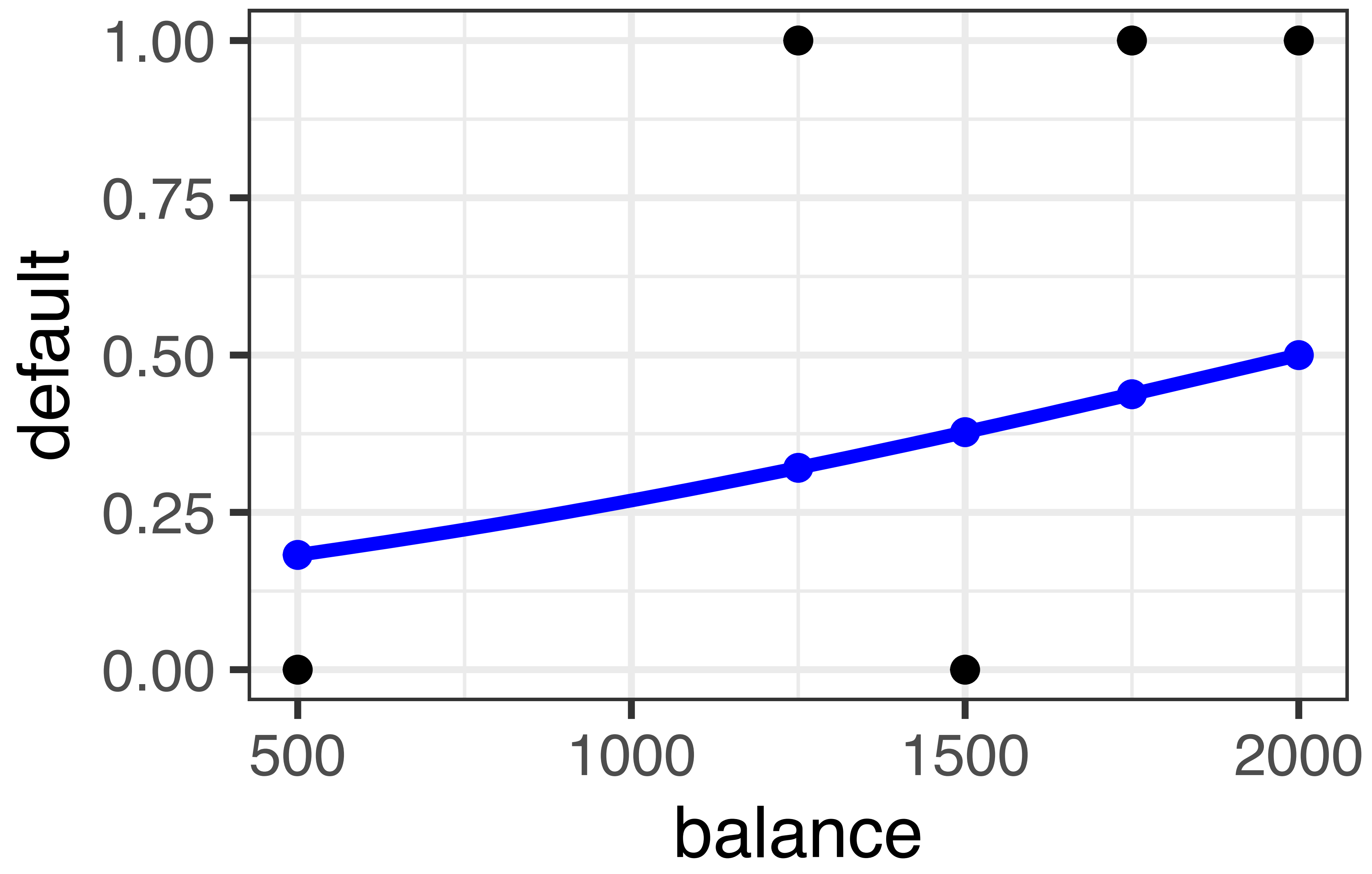
Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:



β_0	β_1
-2.0	0.001

Maximum likelihood estimation

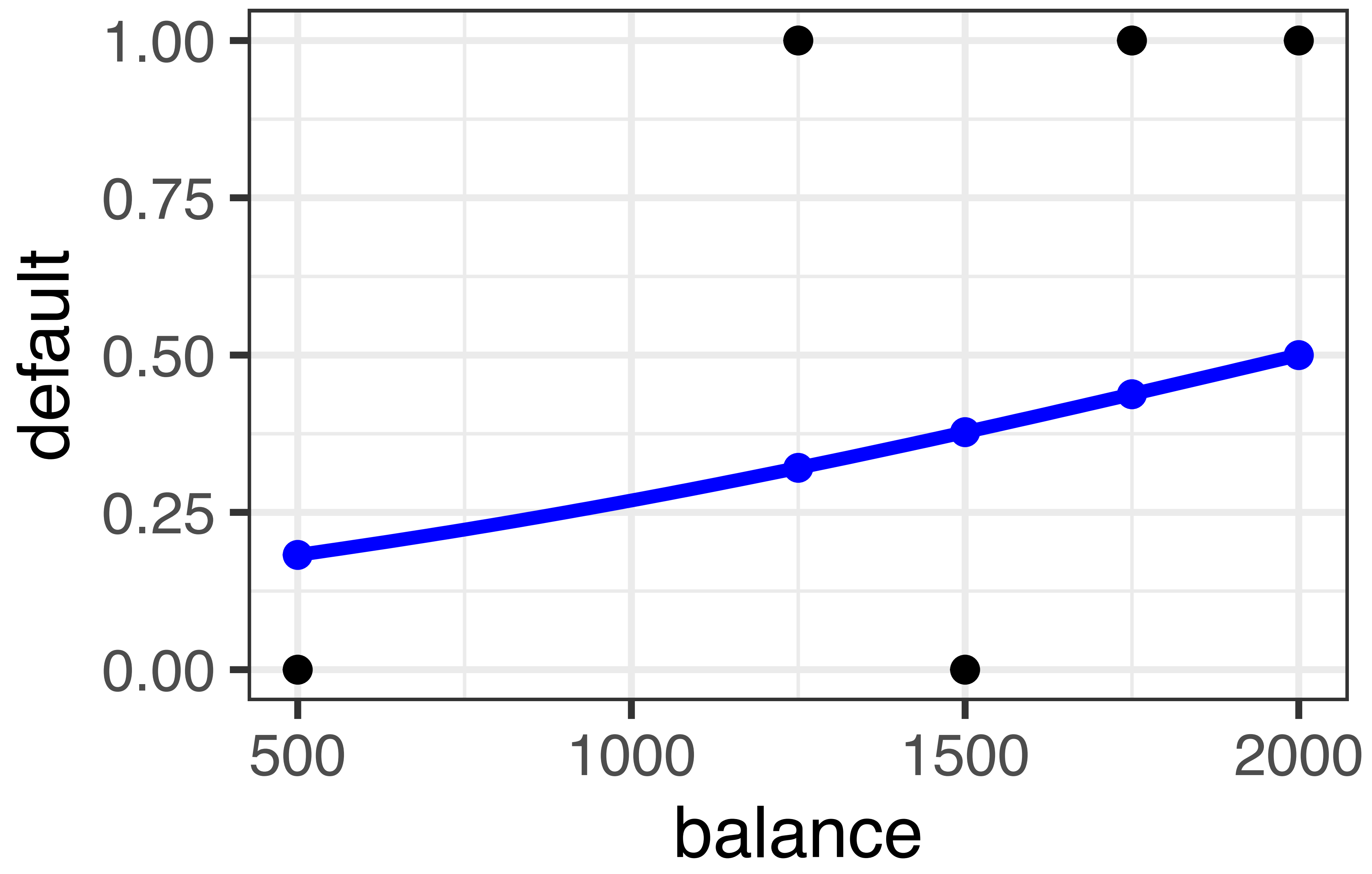
Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:



β_0	β_1	Predicted probabilities
-2.0	0.001	

Maximum likelihood estimation

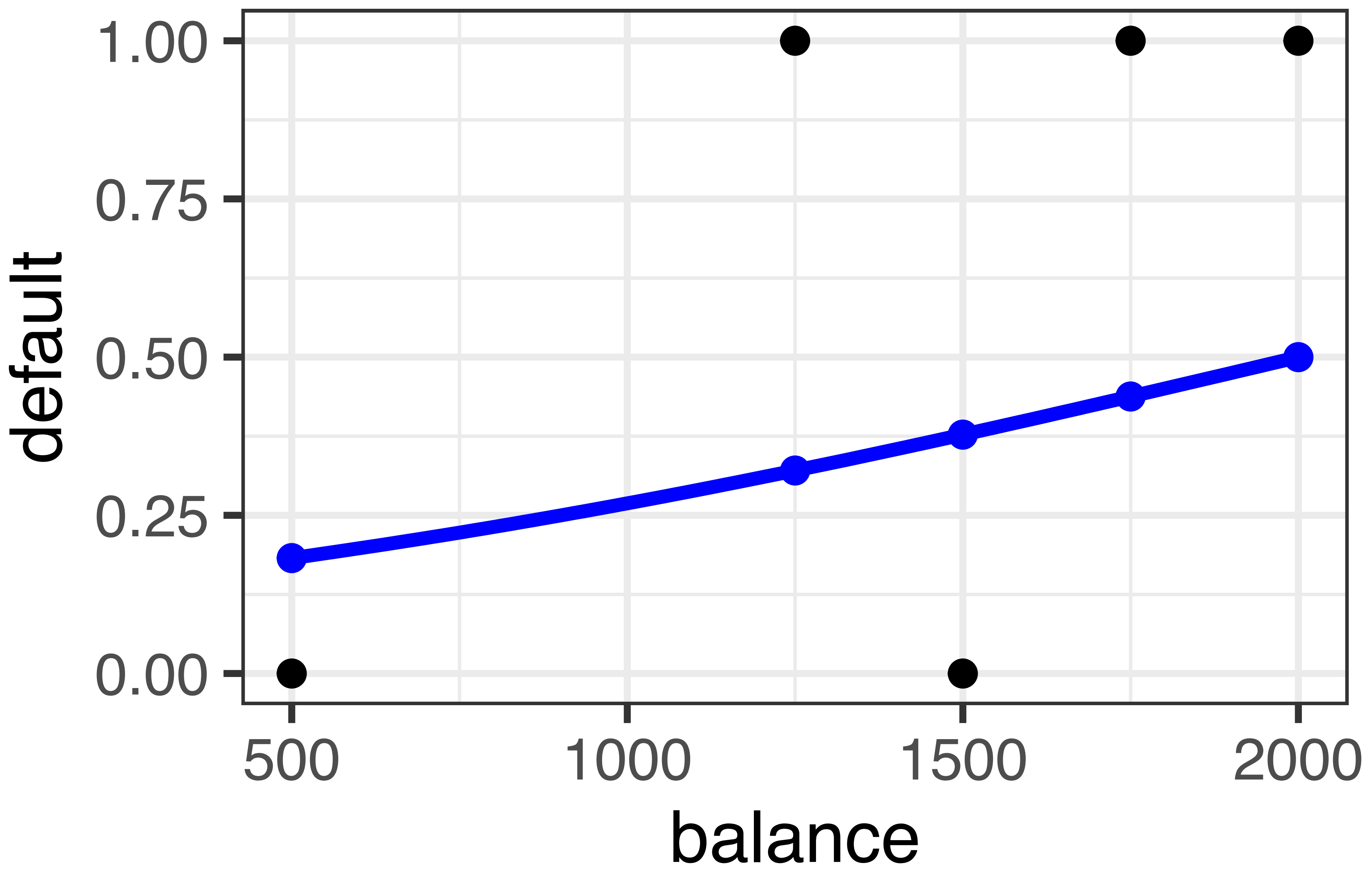
Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:



β_0	β_1	Predicted probabilities
-2.0	0.001	0.8

Maximum likelihood estimation

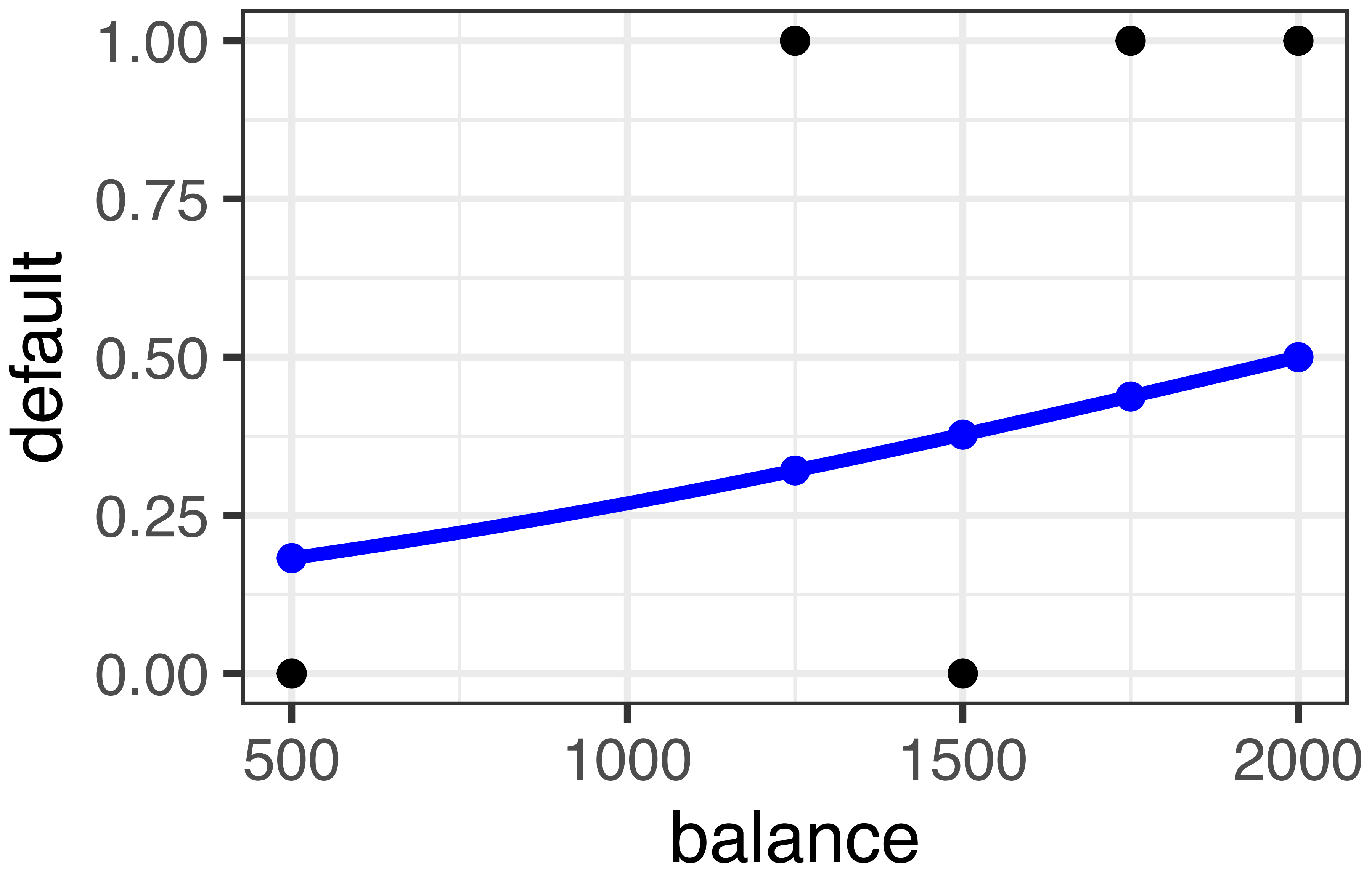
Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:



β_0	β_1	Predicted probabilities				
-2.0	0.001	0.8	0.3			

Maximum likelihood estimation

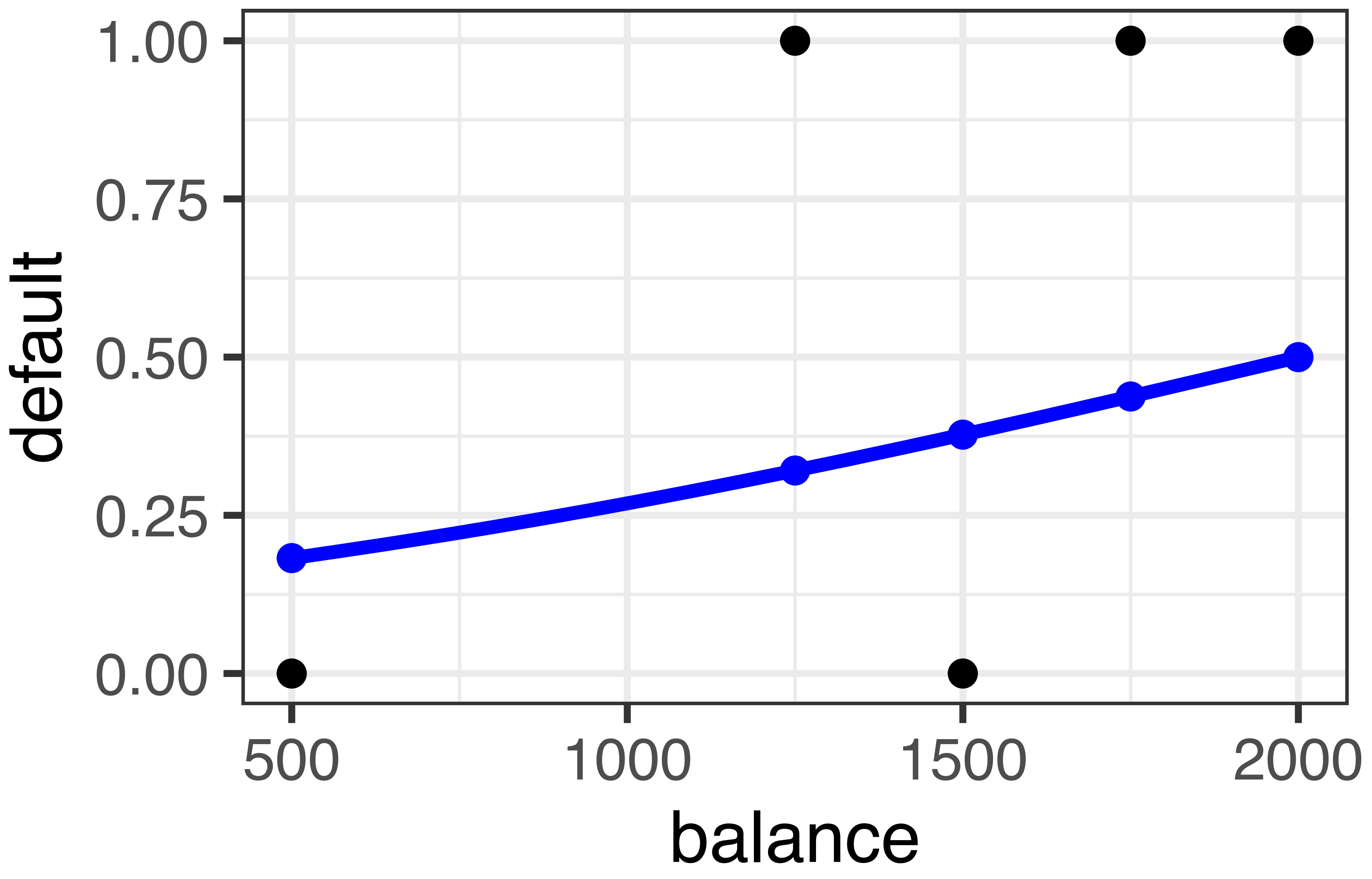
Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:



β_0	β_1	Predicted probabilities				
-2.0	0.001	0.8	0.3	0.6		

Maximum likelihood estimation

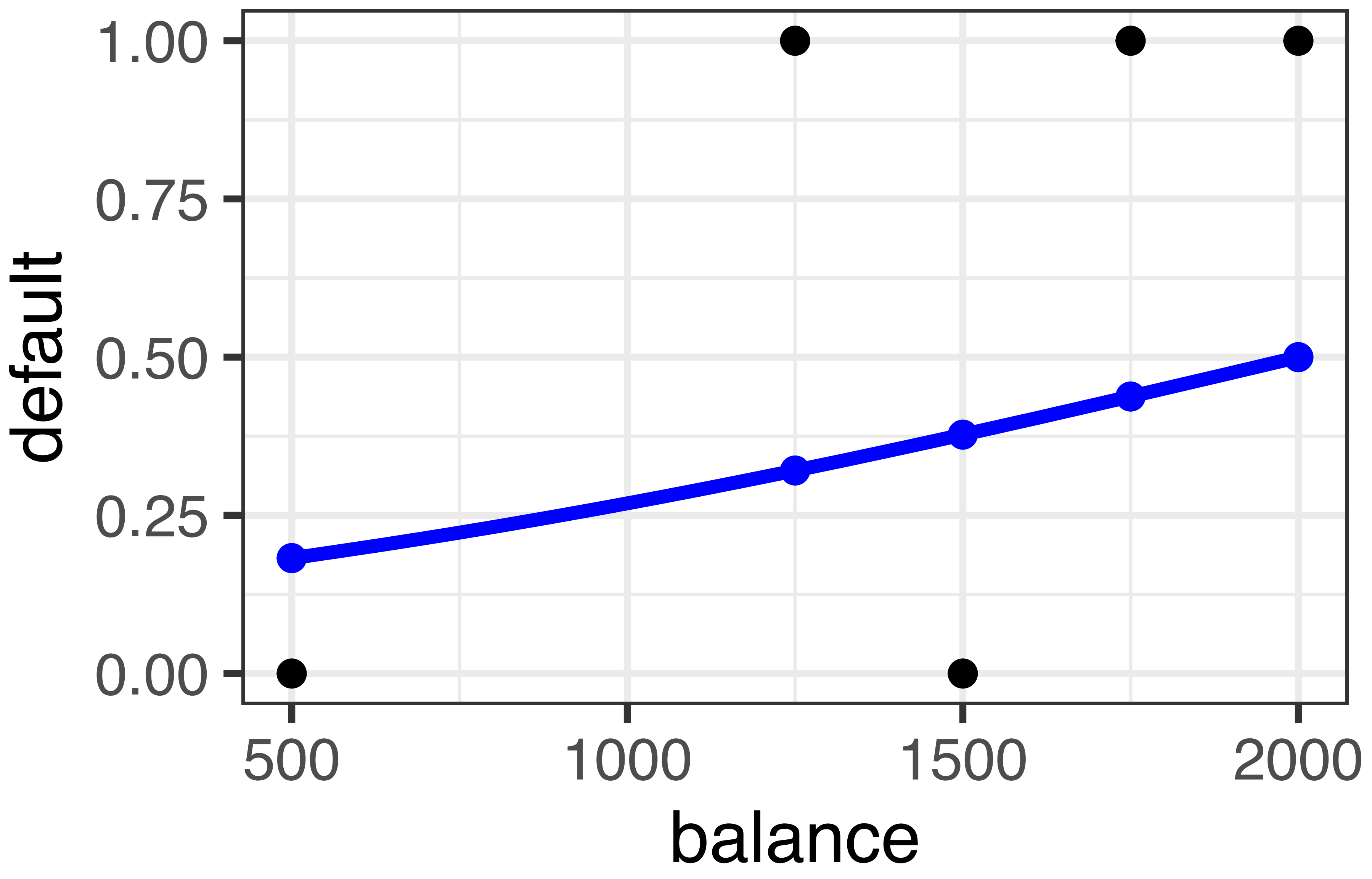
Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:



β_0	β_1	Predicted probabilities				
-2.0	0.001	0.8	0.3	0.6	0.4	0.5

Maximum likelihood estimation

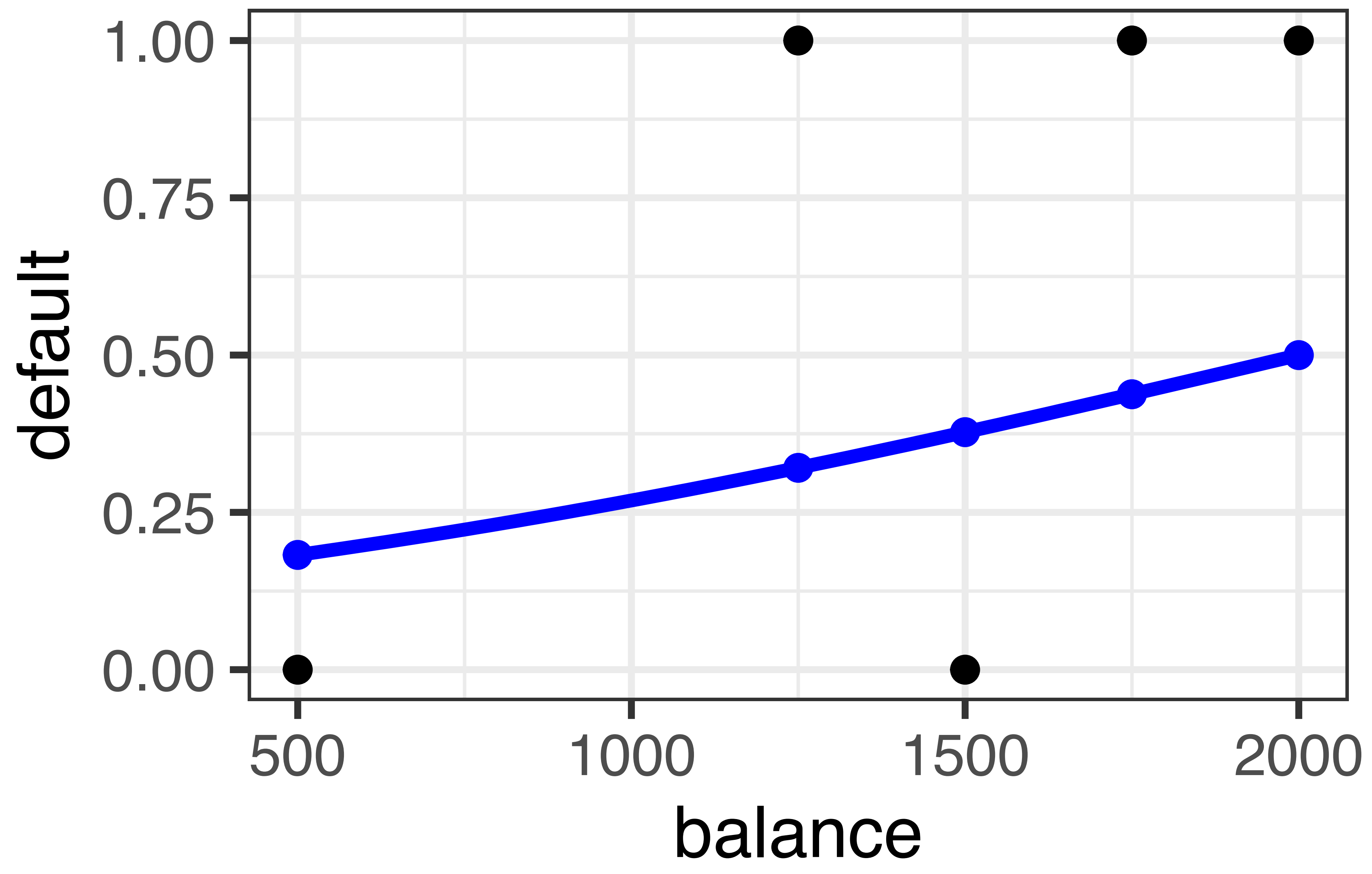
Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:



β_0	β_1	Predicted probabilities					$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	0.3	0.6	0.4	0.5	

Maximum likelihood estimation

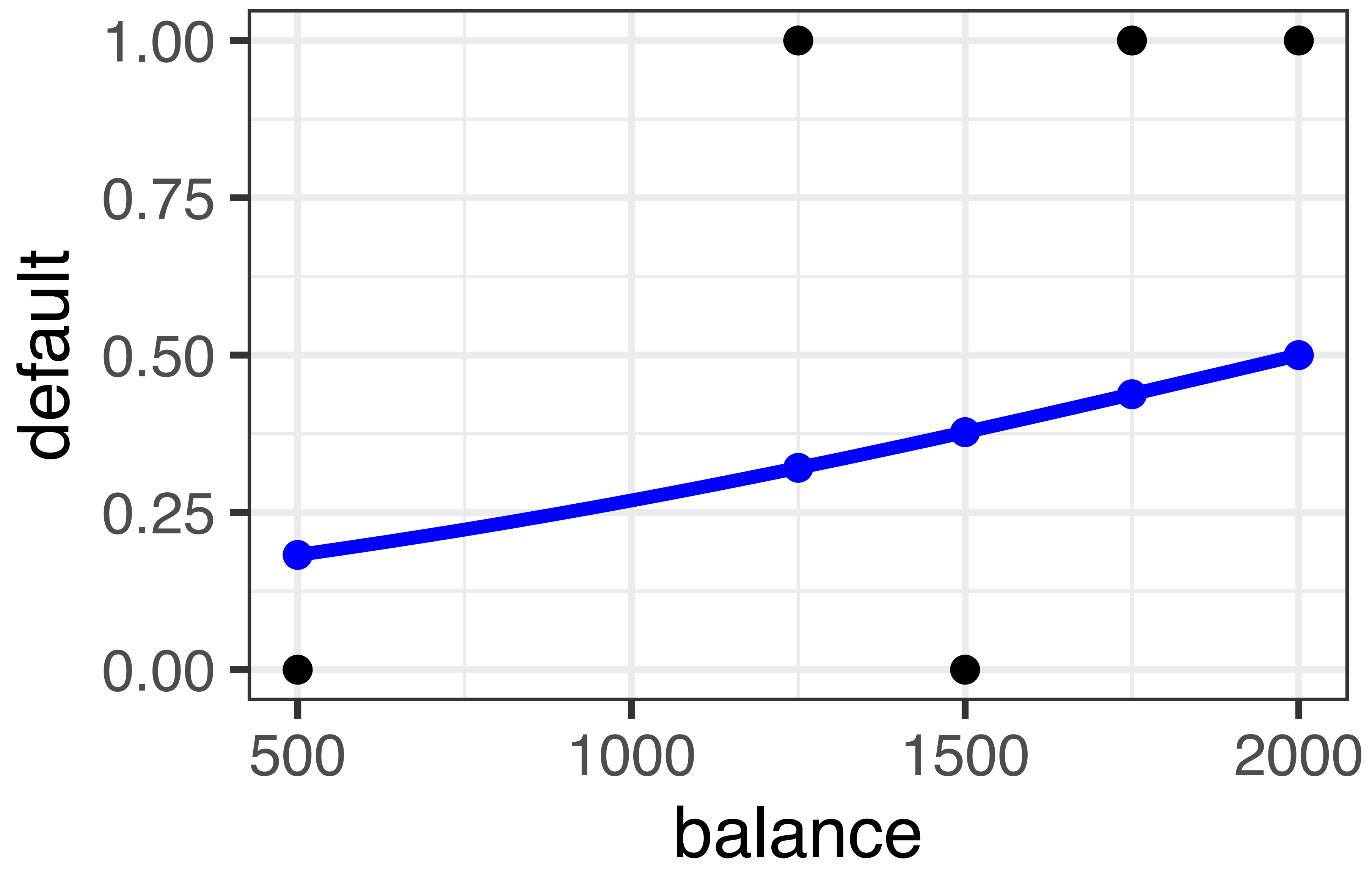
Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:



β_0	β_1	Predicted probabilities	$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8 $\times$ 0.3×0.6×0.4×0.5	

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

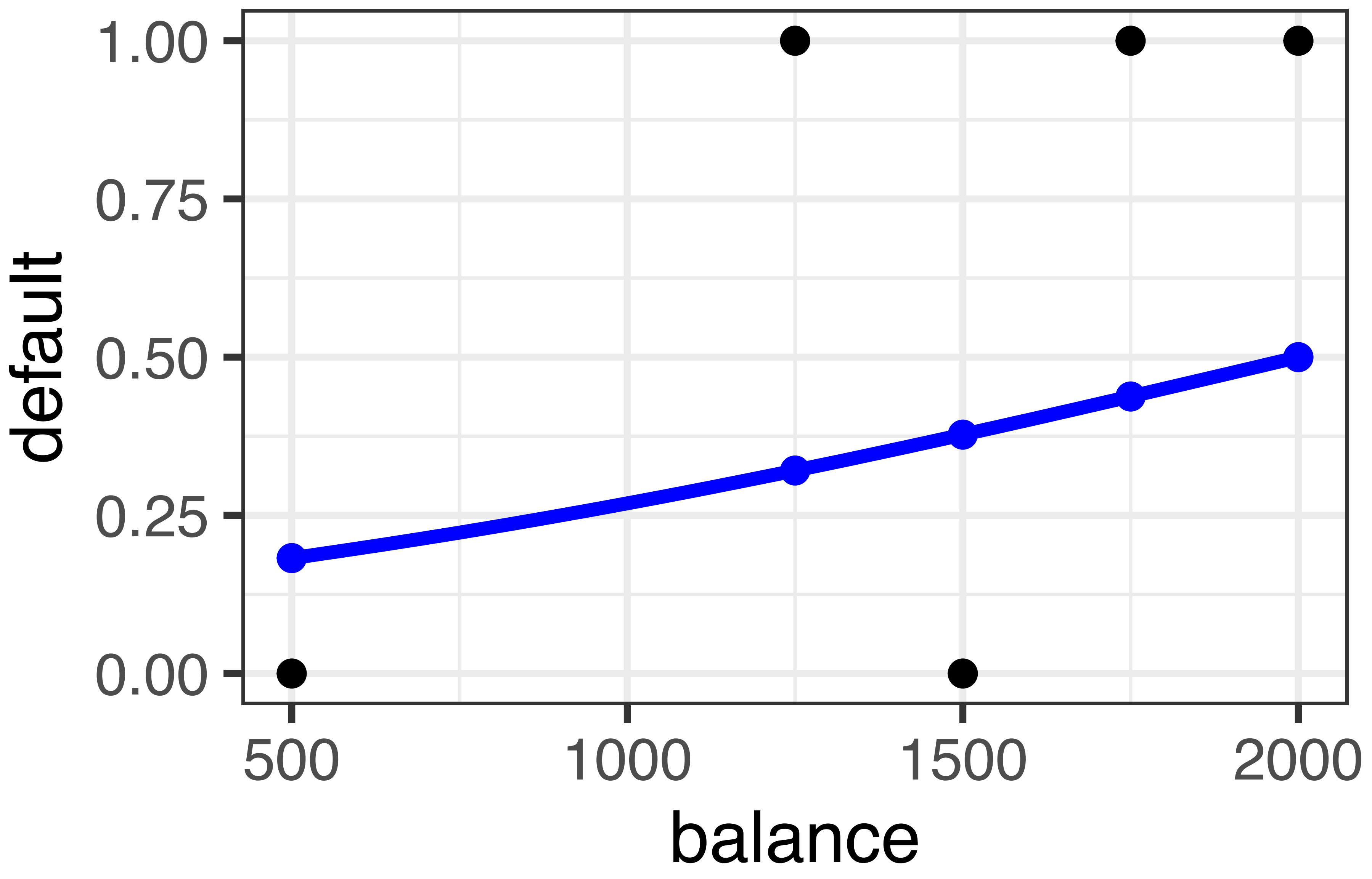


β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.



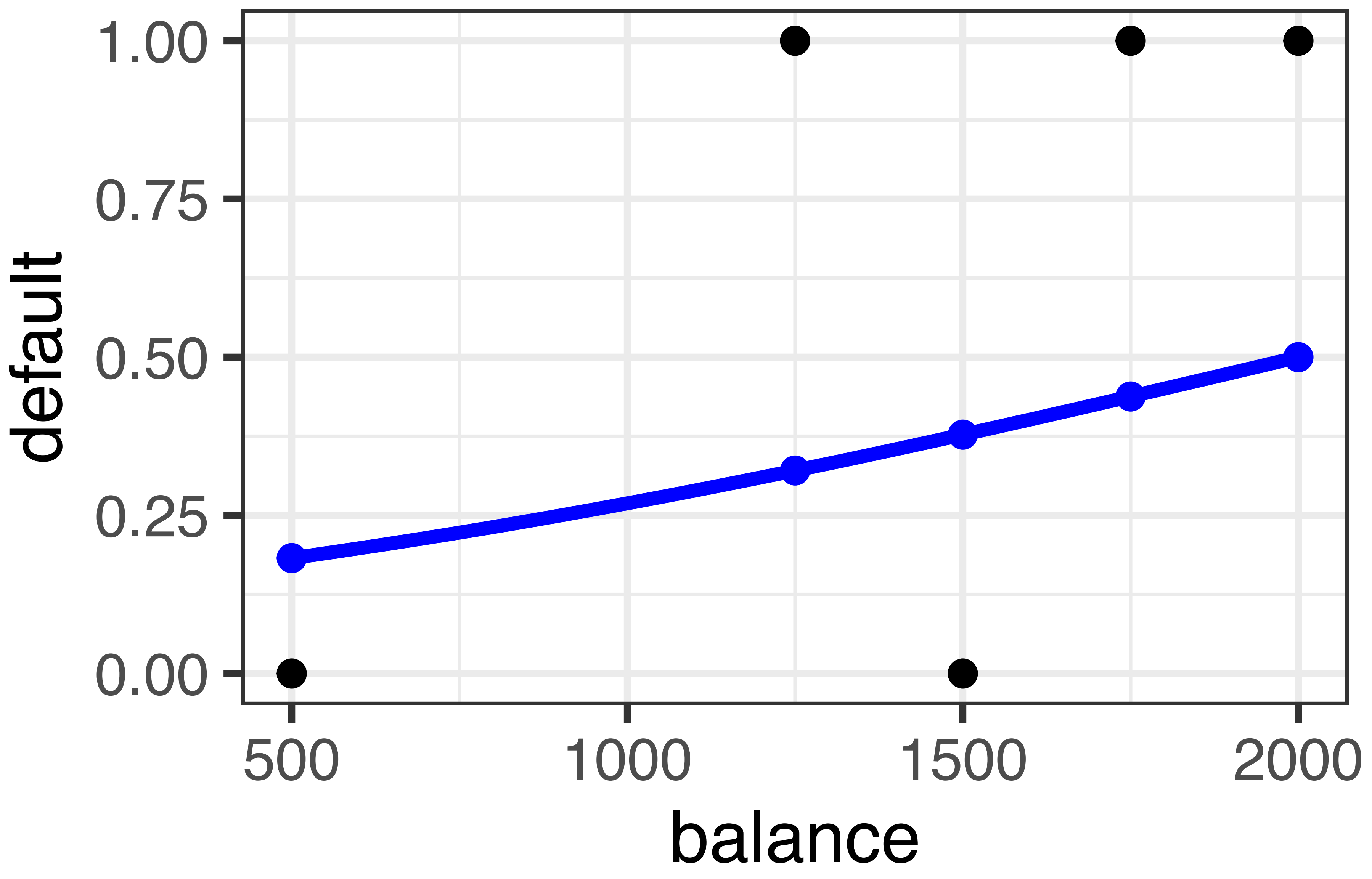
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



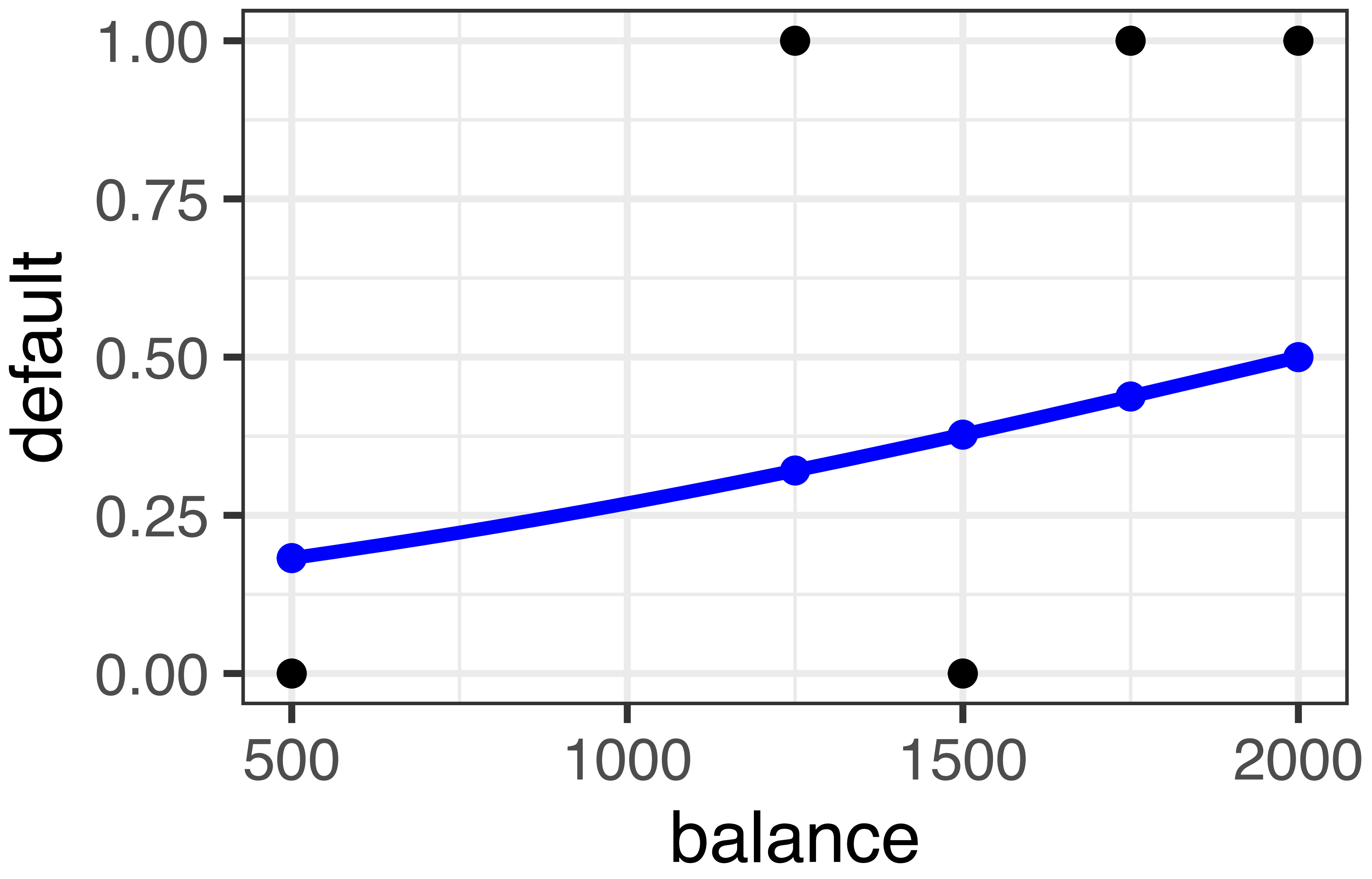
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\hat{\beta}_0, \hat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



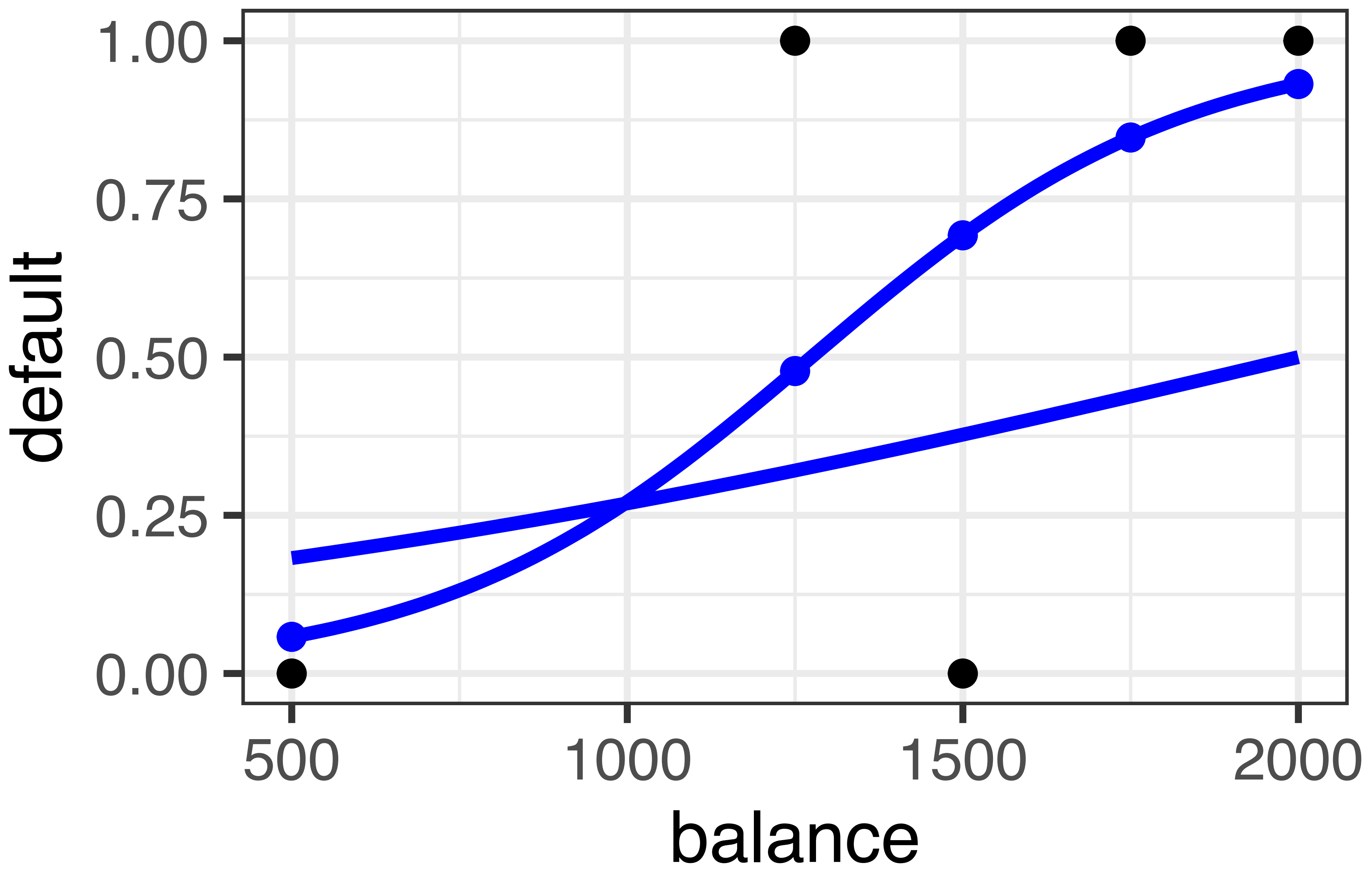
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004				

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



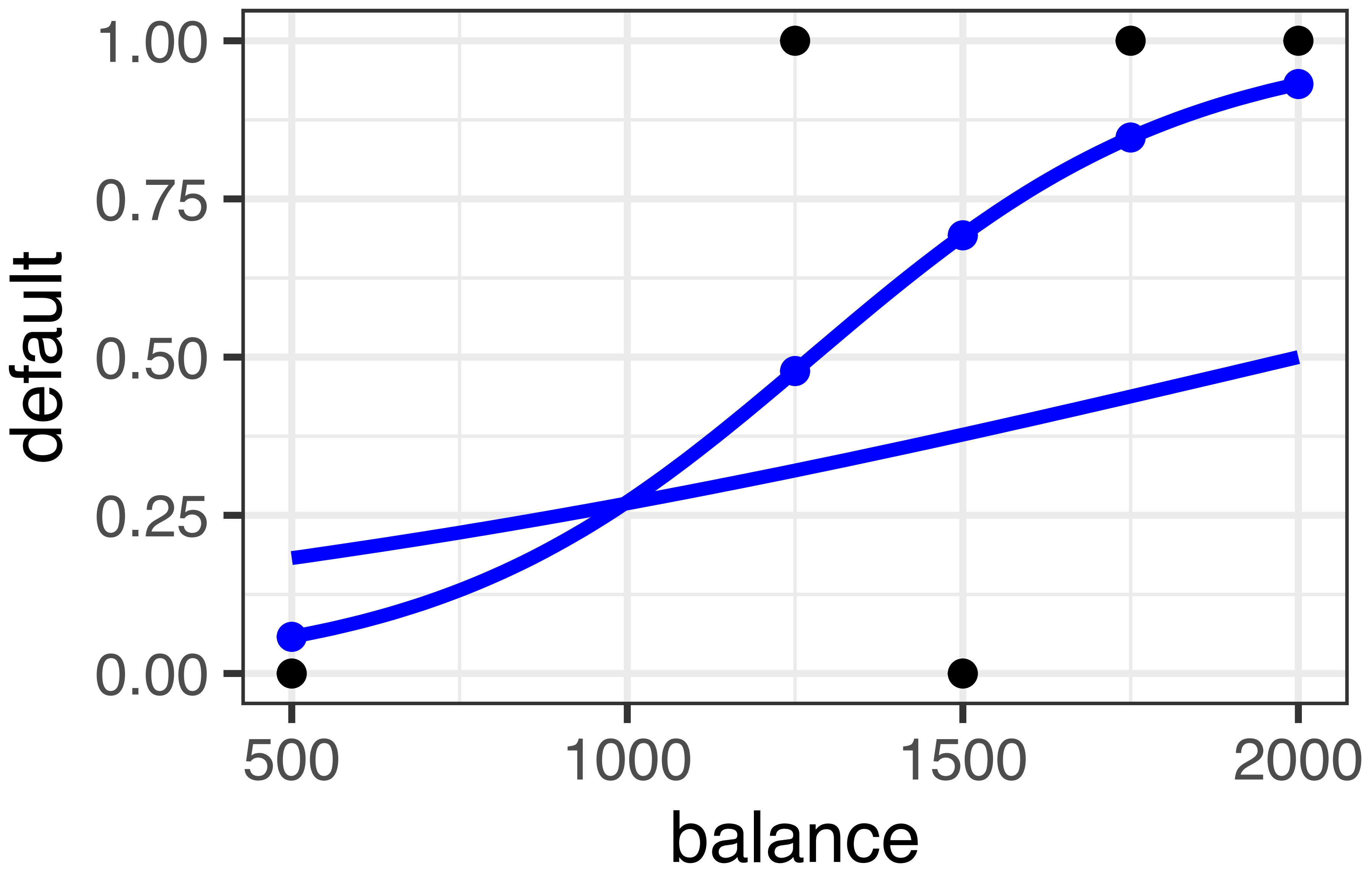
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	$\times$	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004				

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



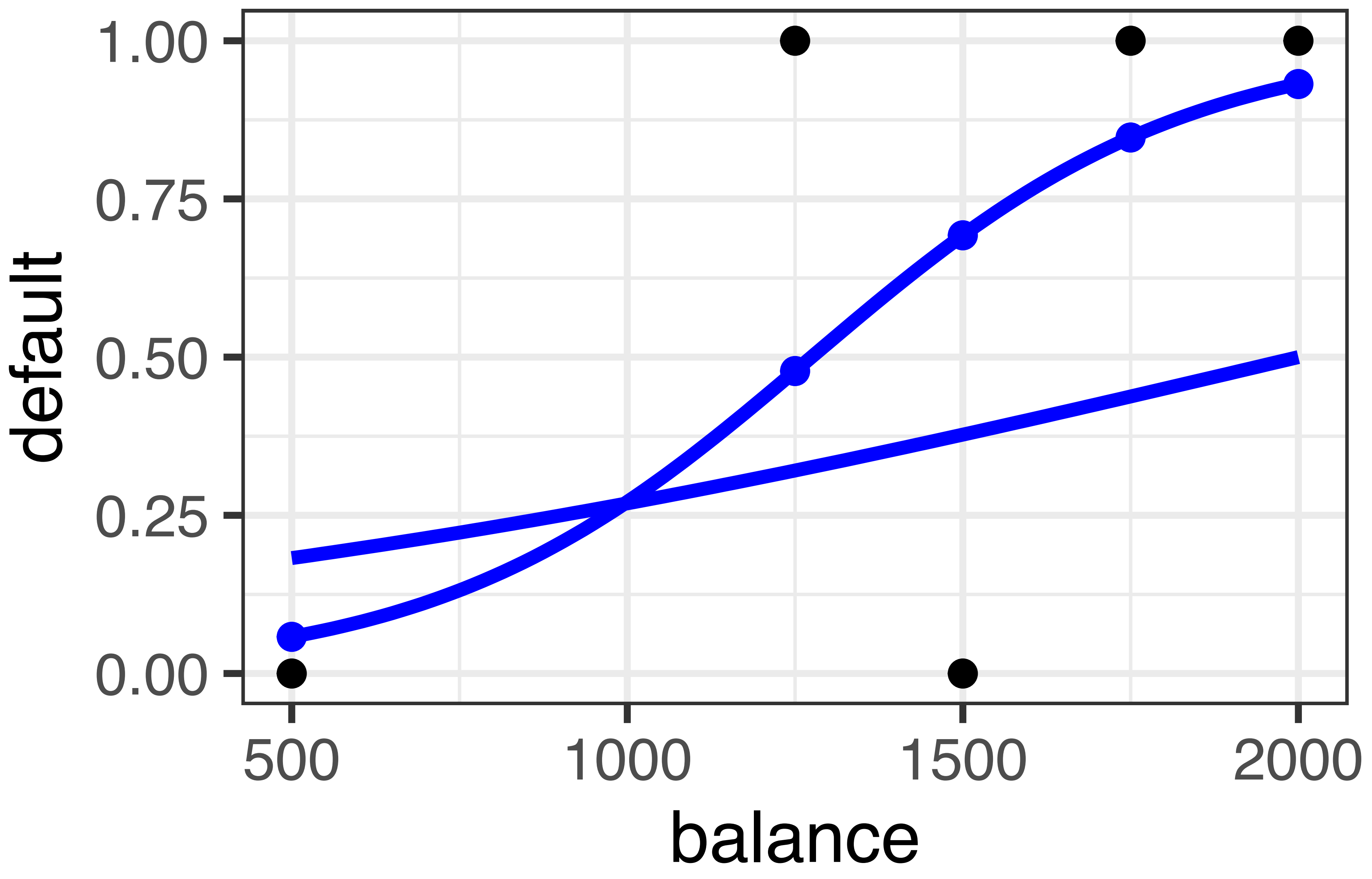
β_0	β_1	Predicted probabilities						$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×	0.6×	0.4×	0.5	= 0.03
-4.6	0.004	0.9		0.5	0.3	0.8	0.9	

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



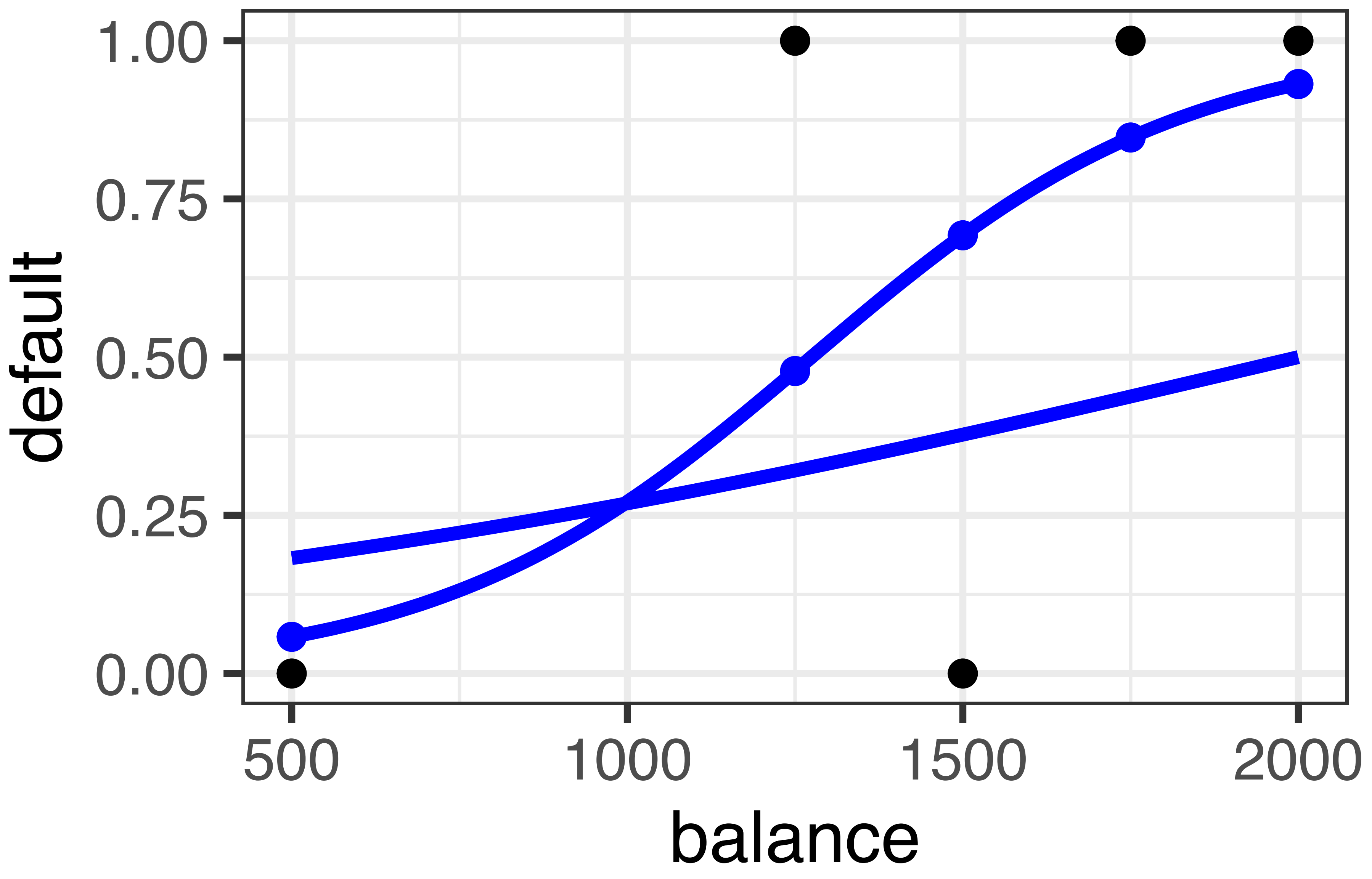
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004	0.9	×	0.5×0.3×0.8×0.9	

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



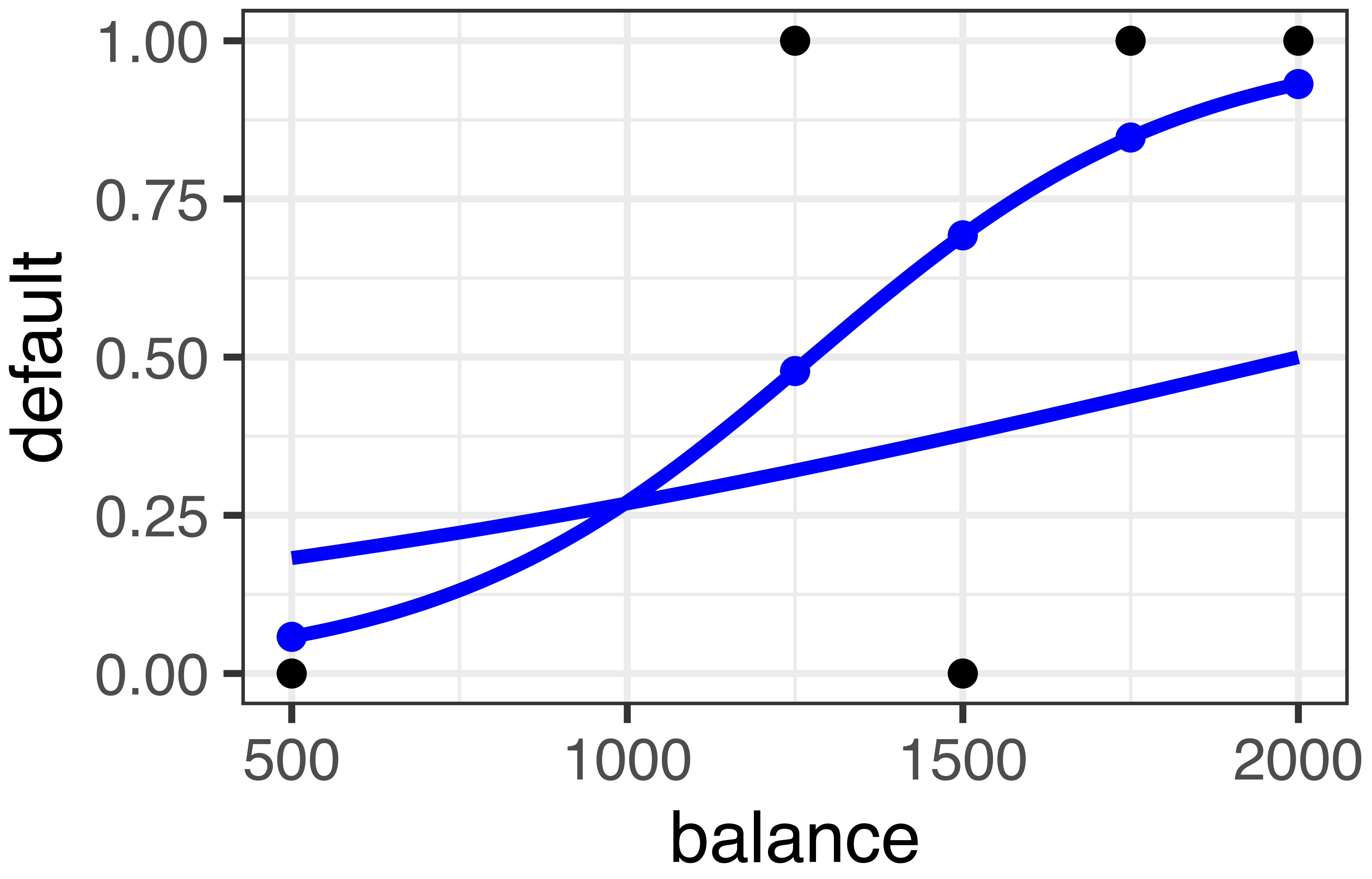
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004	0.9	×	0.5×0.3×0.8×0.9	= 0.1

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



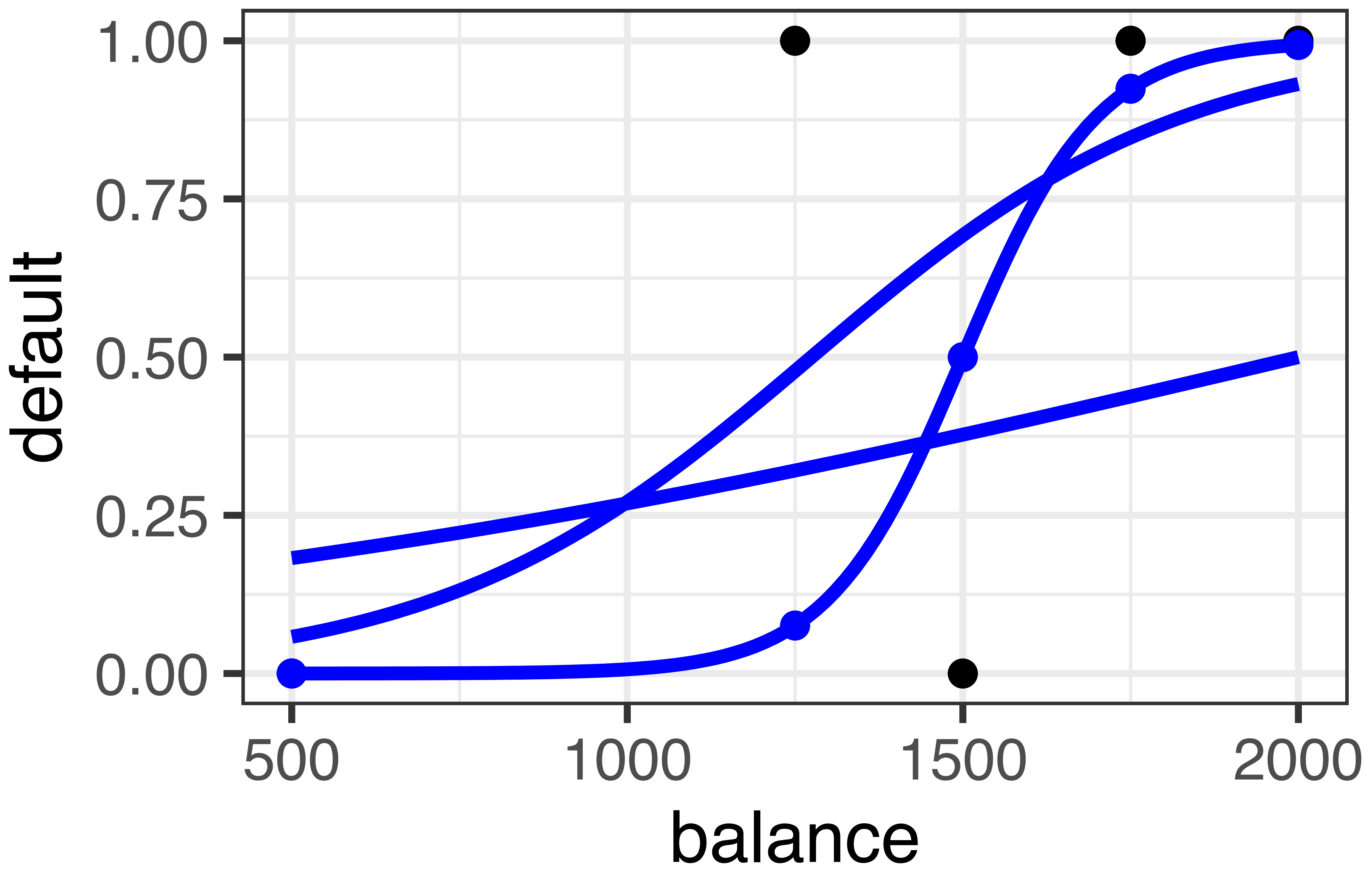
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004	0.9	×	0.5×0.3×0.8×0.9	= 0.1
-15.0	0.01				

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



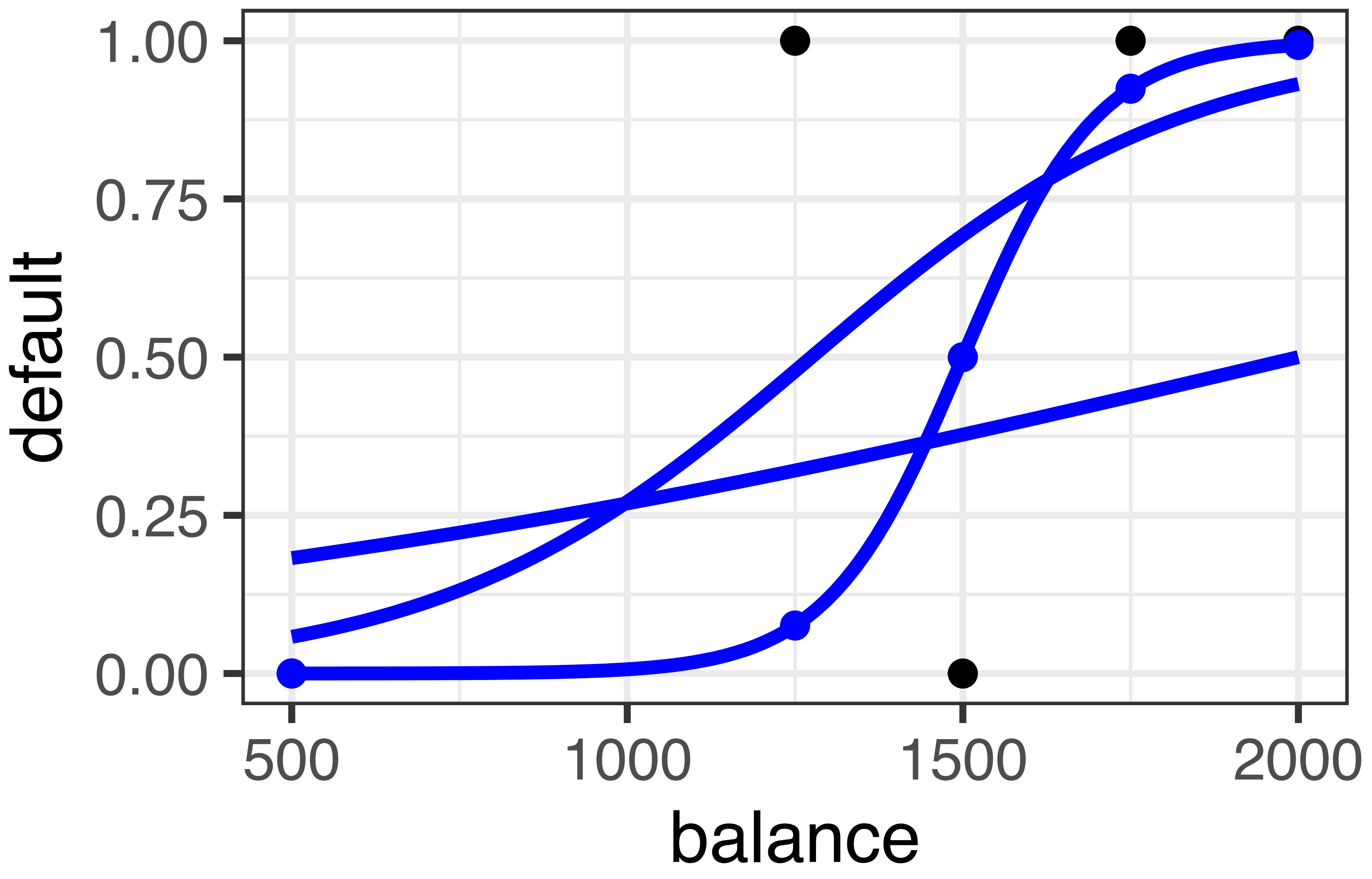
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004	0.9	×	0.5×0.3×0.8×0.9	= 0.1
-15.0	0.01				

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



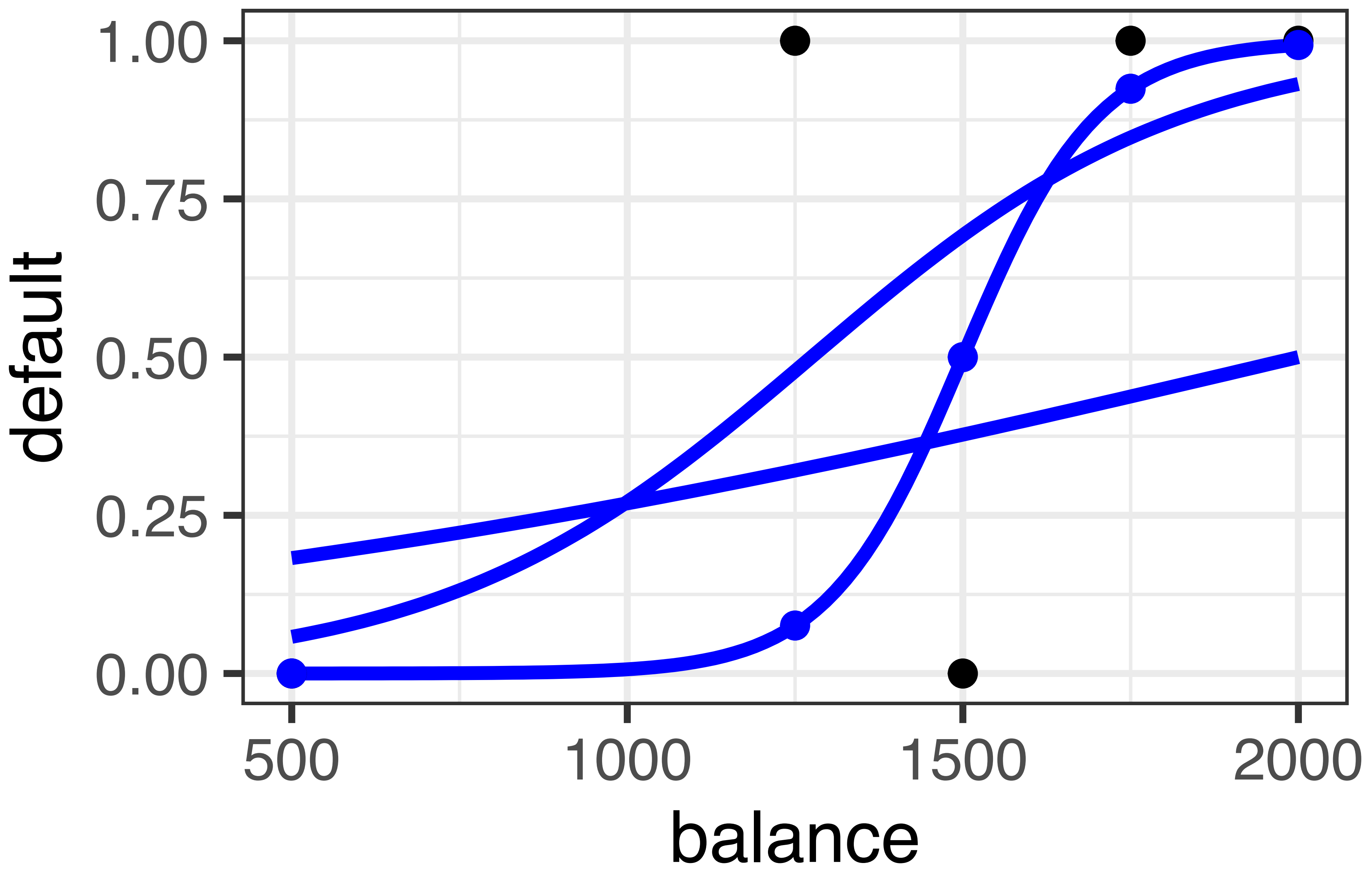
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004	0.9	×	0.5×0.3×0.8×0.9	= 0.1
-15.0	0.01	1.0		0.1 0.5 0.9 1.0	

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



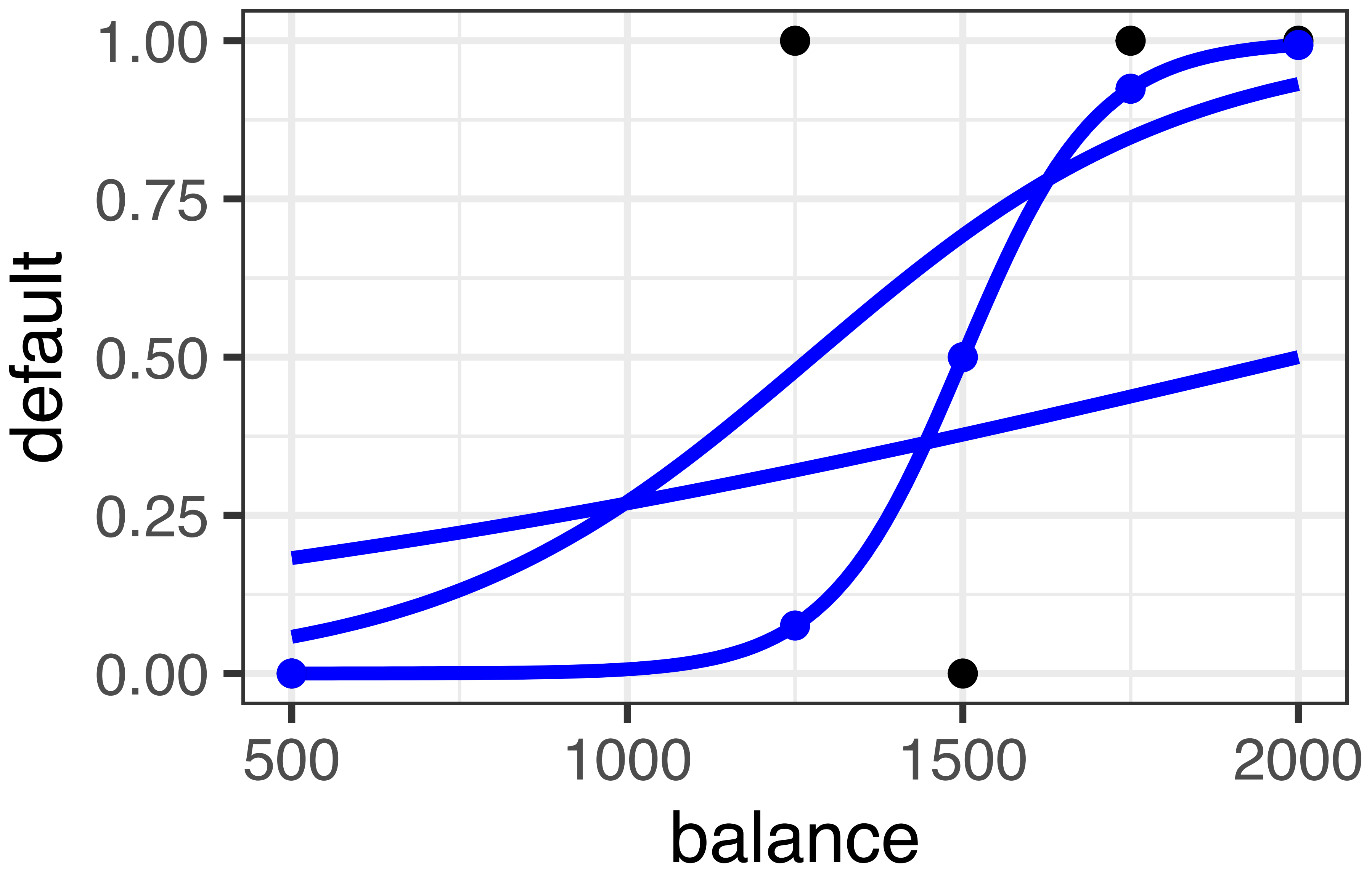
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004	0.9	×	0.5×0.3×0.8×0.9	= 0.1
-15.0	0.01	1.0	×	0.1×0.5×0.9×1.0	

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



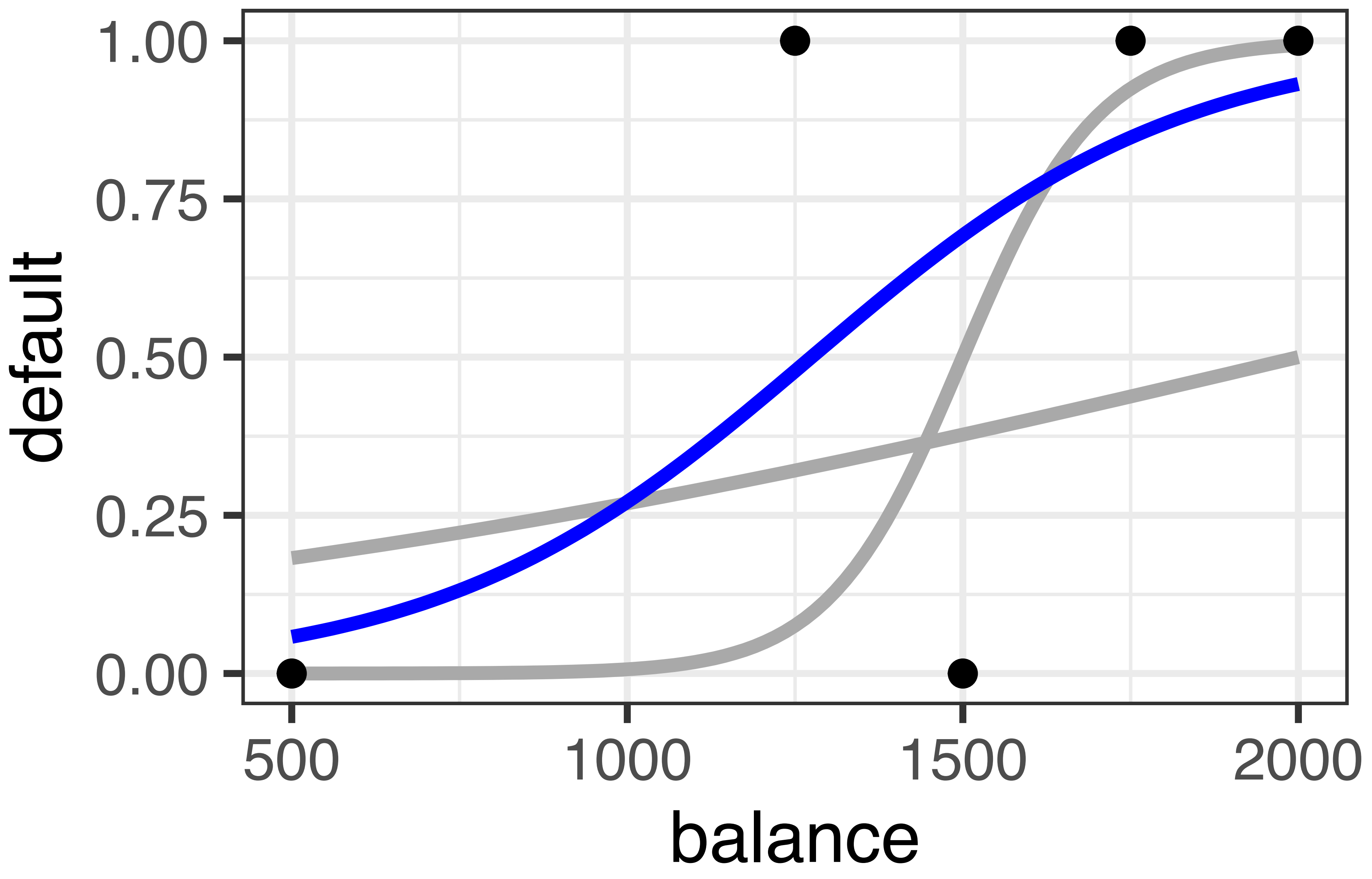
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004	0.9	×	0.5×0.3×0.8×0.9	= 0.1
-15.0	0.01	1.0	×	0.1×0.5×0.9×1.0	= 0.05

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



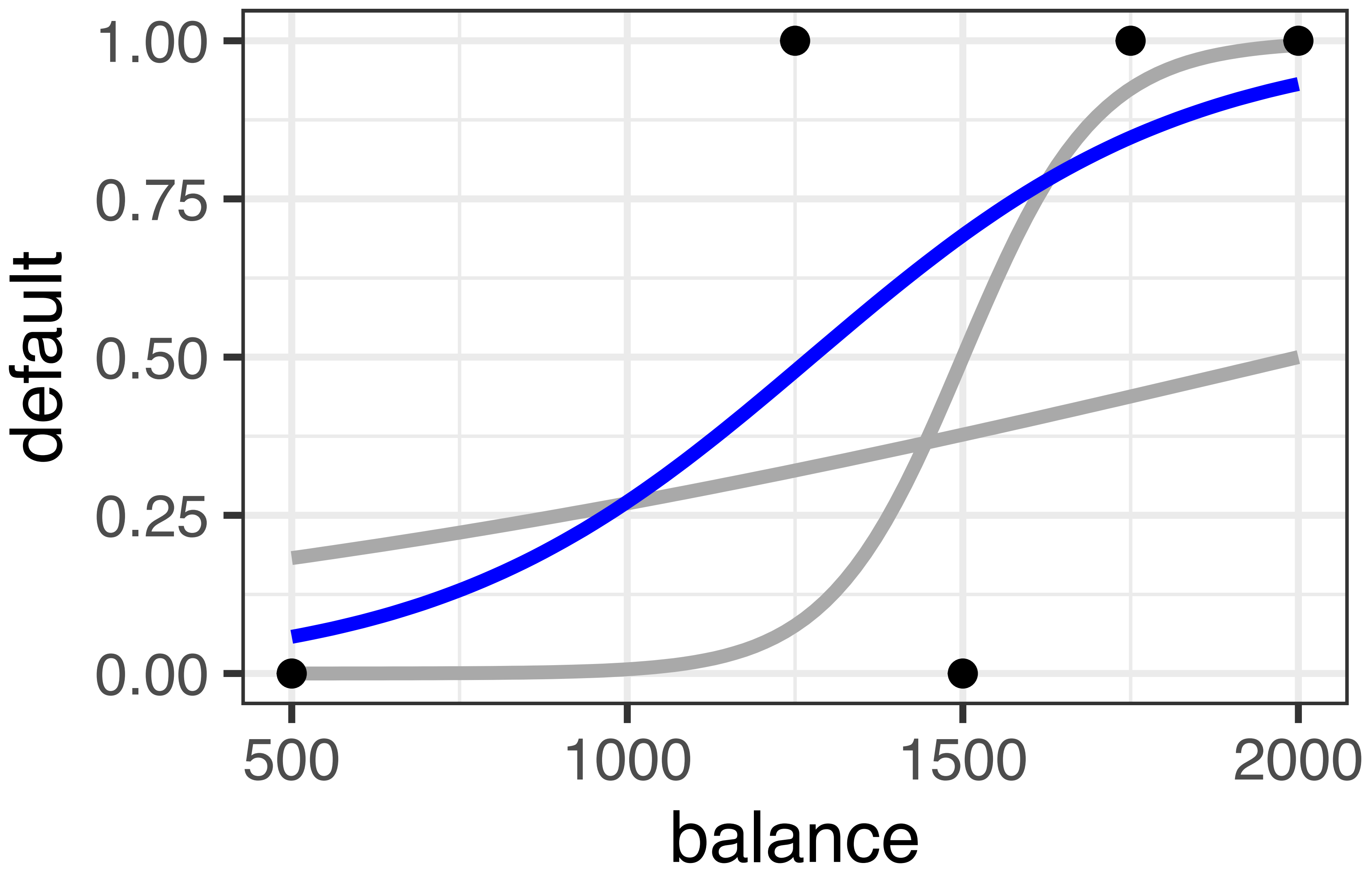
β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
-4.6	0.004	0.9	×	0.5×0.3×0.8×0.9	= 0.1
-15.0	0.01	1.0	×	0.1×0.5×0.9×1.0	= 0.05

Maximum likelihood estimation

Given candidate parameters (β_0, β_1), we define the likelihood $\mathscr{L}(\beta_0, \beta_1)$ as the probability of observing the data under the corresponding model:

The maximum likelihood estimate (MLE) $(\widehat{\beta}_0, \widehat{\beta}_1)$ is defined as the maximizer of $\mathscr{L}(\beta_0, \beta_1)$.

It cannot be written in closed form; it is found via iterative algorithm.



	β_0	β_1	Predicted probabilities			$\mathscr{L}(\beta_0, \beta_1)$
	-2.0	0.001	0.8	×	0.3×0.6×0.4×0.5	= 0.03
$(\widehat{\beta}_0, \widehat{\beta}_1) =$	(-4.6 ,	0.004)	0.9	×	0.5×0.3×0.8×0.9	= 0.1
	-15.0	0.01	1.0	×	0.1×0.5×0.9×1.0	= 0.05

Multiple logistic regression

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

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

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

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

Interpreting logistic regression coefficients

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

For given (student, balance, income),
suppose $\mathbb{P}[\text{default}] = 1/4$.

Interpreting logistic regression coefficients

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

$$\Downarrow$$

For given (student, balance, income), suppose $\mathbb{P}[\text{default}] = 1/4$.

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

Interpreting logistic regression coefficients

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

$$\Downarrow$$

For given (student, balance, income), suppose $\mathbb{P}[\text{default}] = 1/4$.

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

log-odds (the score from before)

Interpreting logistic regression coefficients

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

$\Downarrow$ For given (student, balance, income), suppose $\mathbb{P}[\text{default}] = 1/4$.

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

log-odds (the score from before)

Interpreting logistic regression coefficients

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

$\Downarrow$ For given (student, balance, income), suppose $\mathbb{P}[\text{default}] = 1/4$.

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

log-odds (the score from before)

Then, odds = 1:3 = 1/3 and log-odds = log(1/3) $\approx$ -1.

Interpreting logistic regression coefficients

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

$\Downarrow$ For given (student, balance, income), suppose $\mathbb{P}[\text{default}] = 1/4$.

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

log-odds (the score from before)

$\Downarrow$ Then, odds = 1:3 = 1/3 and log-odds = log(1/3) ≈ -1.

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

Interpreting logistic regression coefficients

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

$\Downarrow$ For given (student, balance, income), suppose $\mathbb{P}[\text{default}] = 1/4$.

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

log-odds (the score from before)

$\Downarrow$ Then, odds = 1:3 = 1/3 and log-odds = log(1/3) ≈ -1.

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

If $\beta_2 = 1/250$, then increasing balance by $500
Increases log-odds by 2; new log-odds is -1 + 2 =1.

Interpreting logistic regression coefficients

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

$\Downarrow$ For given (student, balance, income), suppose $\mathbb{P}[\text{default}] = 1/4$.

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

log-odds (the score from before)

$\Downarrow$ Then, odds = 1:3 = 1/3 and log-odds = log(1/3) ≈ -1.

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

$\Downarrow$ If $\beta_2 = 1/250$, then increasing balance by \$500 Increases log-odds by 2; new log-odds is -1 + 2 =1.

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

Interpreting logistic regression coefficients

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

$\Downarrow$ For given (student, balance, income), suppose $\mathbb{P}[\text{default}] = 1/4$.

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

log-odds (the score from before)

$\Downarrow$ Then, odds = 1:3 = 1/3 and log-odds = log(1/3) $\approx$ -1.

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

$\Downarrow$ If $\beta_2 = 1/250$, then increasing balance by \$500 Increases log-odds by 2; new log-odds is -1 + 2 =1.

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

New odds are $e^1 \approx 2.7 = 2.7 : 1$, so new prob is 2.7/3.7 $\approx$ 0.7.
Odds went from e^{-1} (1/3) to e^1 (2.7), increase by factor of $e^2 \approx 7.5$.

Classification via logistic regression

$$\text{default} = \begin{cases} \text{Yes}, & \text{if } \widehat{\mathbb{P}}[\text{default}] \geq 0.5; \\ \text{No}, & \text{if } \widehat{\mathbb{P}}[\text{default}] < 0.5. \end{cases}$$

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

Classification via logistic regression

$$\text{default} = \begin{cases} \text{Yes}, & \text{if } \widehat{\mathbb{P}}[\text{default}] \geq 0.5; \\ \text{No}, & \text{if } \widehat{\mathbb{P}}[\text{default}] < 0.5. \end{cases}$$

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

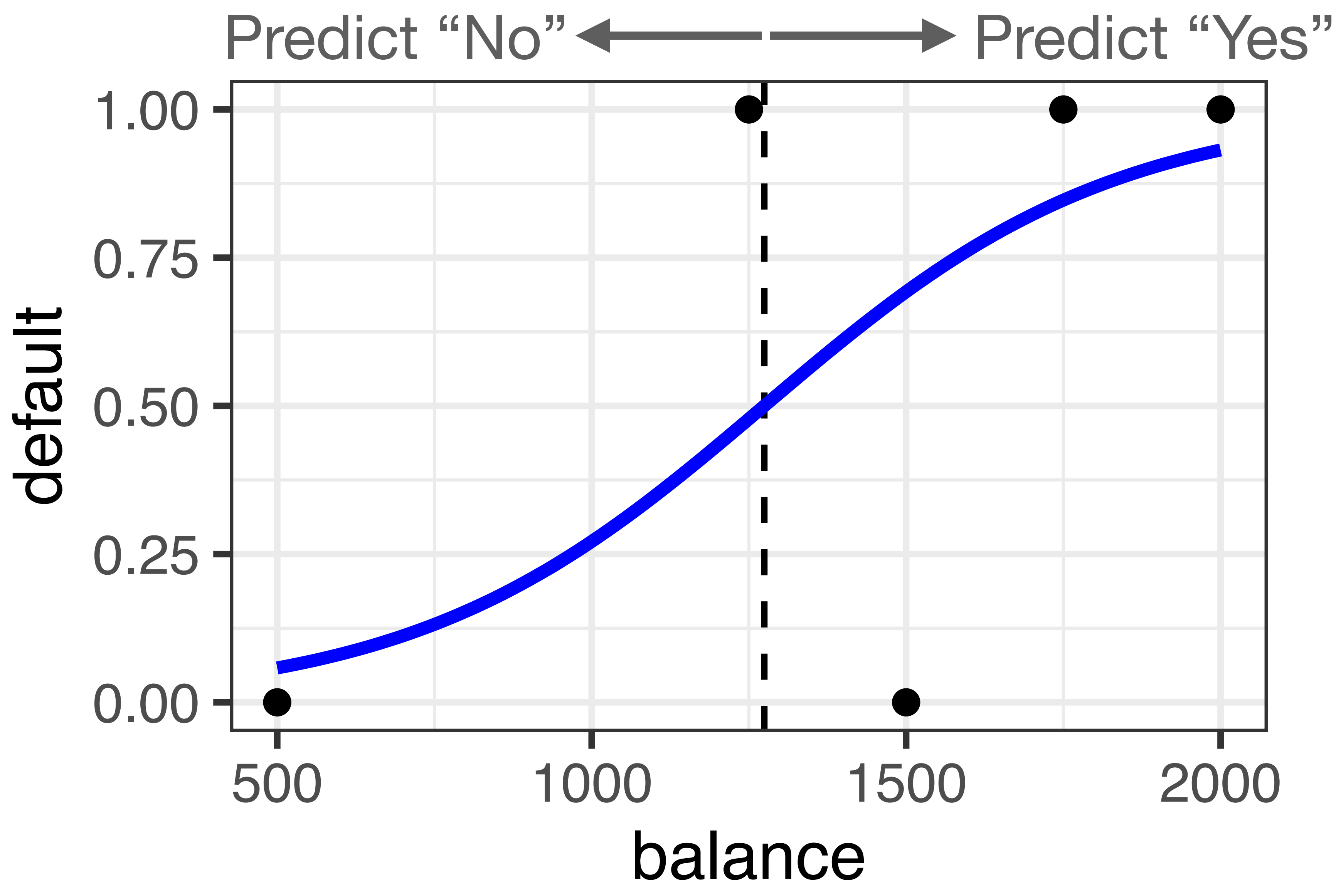
Logistic regression has a linear decision boundary.

Classification via logistic regression

$$\text{default} = \begin{cases} \text{Yes}, & \text{if } \widehat{\mathbb{P}}[\text{default}] \geq 0.5; \\ \text{No}, & \text{if } \widehat{\mathbb{P}}[\text{default}] < 0.5. \end{cases}$$

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

Logistic regression has a linear decision boundary.

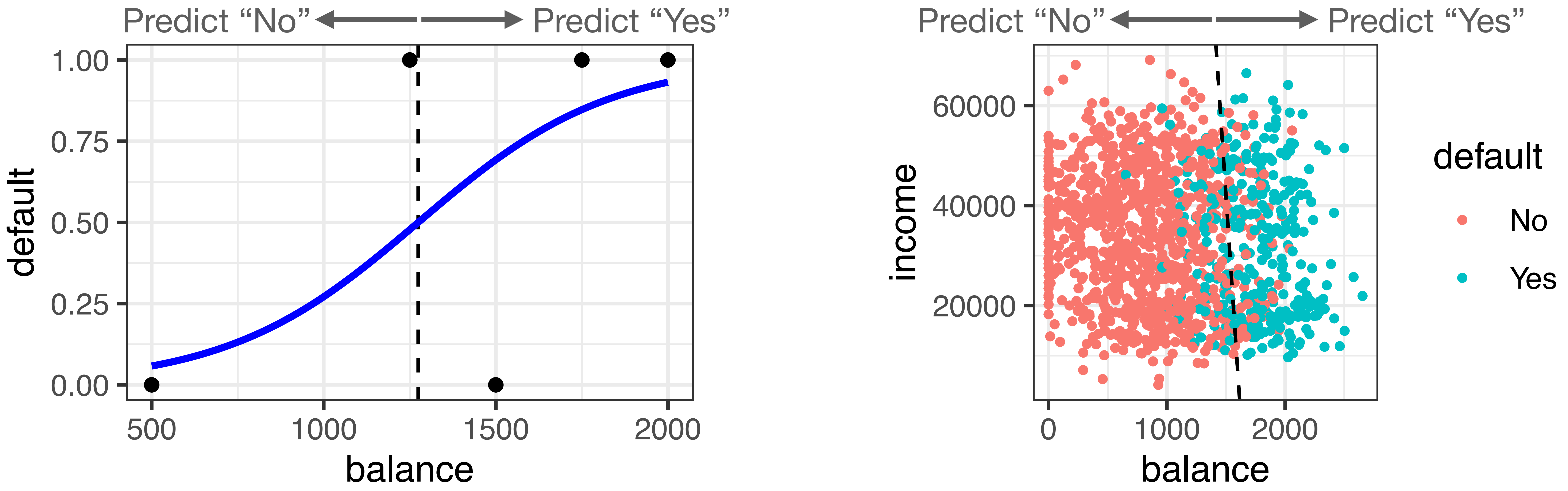


Classification via logistic regression

$$\text{default} = \begin{cases} \text{Yes}, & \text{if } \widehat{\mathbb{P}}[\text{default}] \geq 0.5; \\ \text{No}, & \text{if } \widehat{\mathbb{P}}[\text{default}] < 0.5. \end{cases}$$

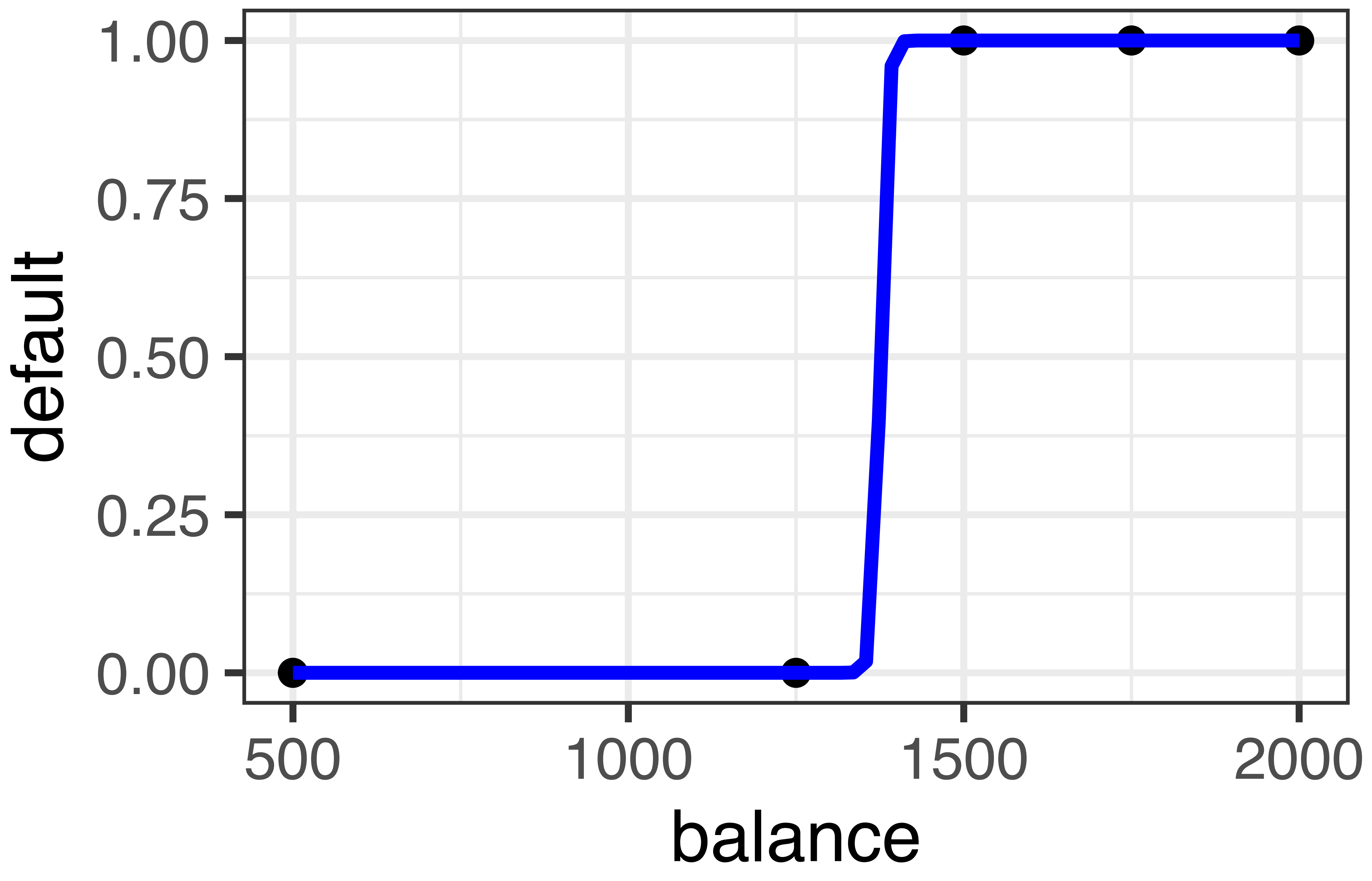
$$\widehat{\mathbb{P}}[\text{default}] > 0.5 \iff \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{student} + \hat{\beta}_2 \cdot \text{balance} + \hat{\beta}_3 \cdot \text{income} > 0$$

Logistic regression has a linear decision boundary.



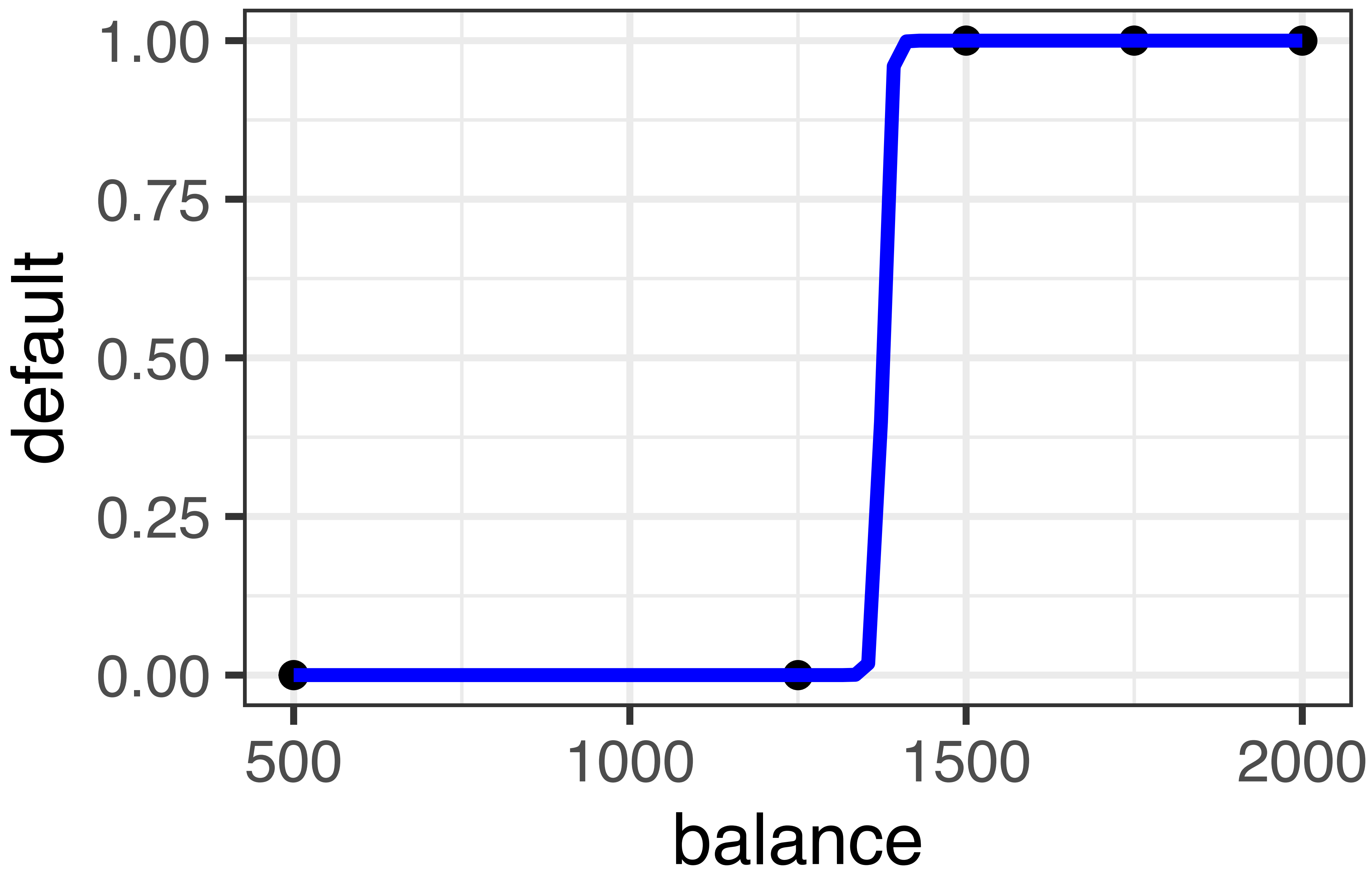
Caution: Separable data

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



Caution: 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.

Summary

Summary

Response type	Continuous	Binary

Summary

Response type	Continuous	Binary
Most common predictive model	Linear regression	Logistic regression

Summary

Response type	Continuous	Binary
Most common predictive model	Linear regression	Logistic regression
Measure of fit	Mean squared error	Likelihood

Summary

Response type	Continuous	Binary
Most common predictive model	Linear regression	Logistic regression
Measure of fit	Mean squared error	Likelihood
Estimating coefficients	Least squares (closed form)	Maximum likelihood (iterative)

Summary

Response type	Continuous	Binary
Most common predictive model	Linear regression	Logistic regression
Measure of fit	Mean squared error	Likelihood
Estimating coefficients	Least squares (closed form)	Maximum likelihood (iterative)
Interpreting coefficients	Unit increase in $X_j \rightarrow$ increase in mean of Y by β_j	Unit increase in $X_j \rightarrow$ increase in odds of Y by e^{β_j}

Summary

Response type	Continuous	Binary
Most common predictive model	Linear regression	Logistic regression
Measure of fit	Mean squared error	Likelihood
Estimating coefficients	Least squares (closed form)	Maximum likelihood (iterative)
Interpreting coefficients	Unit increase in $X_j \rightarrow$ increase in mean of Y by β_j	Unit increase in $X_j \rightarrow$ increase in odds of Y by e^{β_j}

Quiz practice