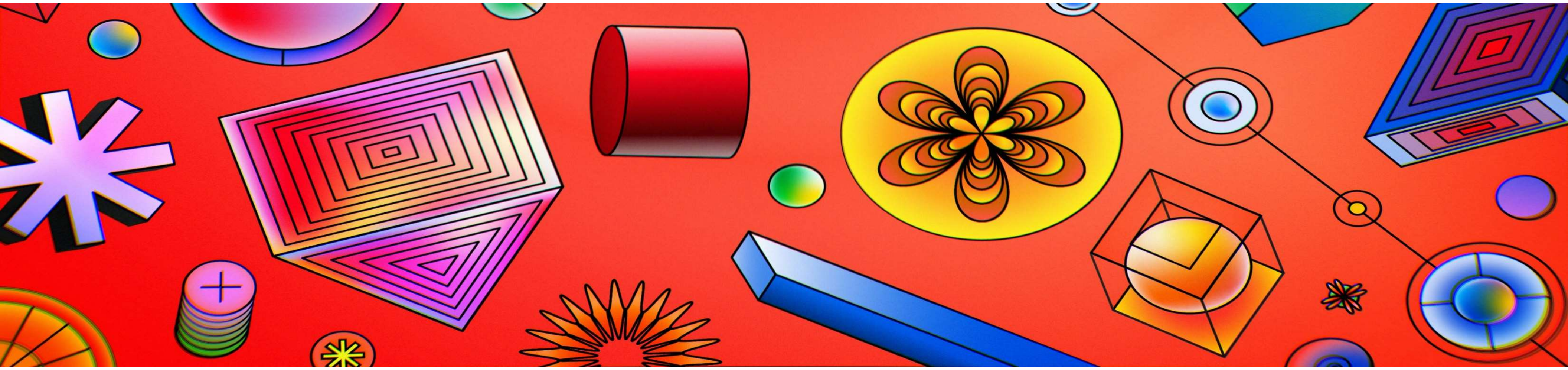
Fall 2023

BIF524/CSC463 Data Mining

Classification

Eileen Marie Hanna, *PhD*

10/10/2023

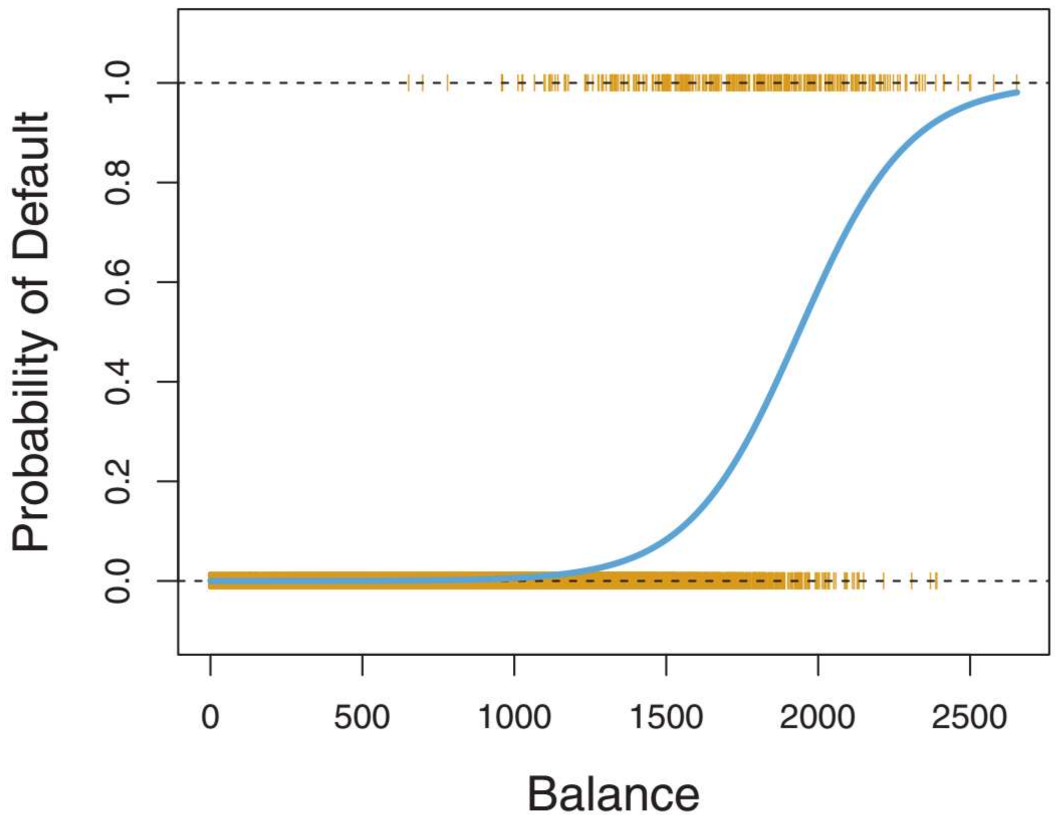


Logistic regression

- The produced curve will always have an **S-shape** depicting a sensible prediction.

- Such model **can capture more of the range of probabilities** than the linear model.

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



$X_1, X_2, \dots, X_n$ joint density function

$$f(x_1, x_2, \dots, x_n | \theta)$$

Given $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$, find function of θ:

$$\ell(\theta) = \ell(\theta | x_1, x_2, \dots, x_n)$$

$$\ell(\theta_1 | x) > \ell(\theta_2 | x)$$

Logistic function

odds
between 0 and ∞

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

- e.g., if 2 out of 10 persons default -> $p(X) = 0.2$ -> odds $= 0.25$
  - if 8 out of 10 persons default -> $p(X) = 0.8$ -> odds $= 4$
    - Values close to 0 correspond to very low probabilities of default.
    - Those close to ∞ correspond to very high probabilities of default.

Logistic function

odds
between 0 and ∞

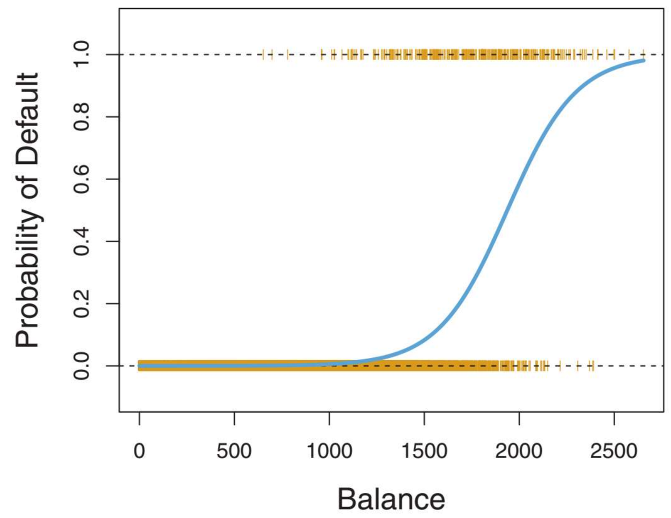
$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \longrightarrow \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

- e.g., if 2 out of 10 persons default -> $p(X) = 0.2$ -> odds = 0.25
  - if 8 out of 10 persons default -> $p(X) = 0.8$ -> odds = 4

By taking the logarithmic on both sides:

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

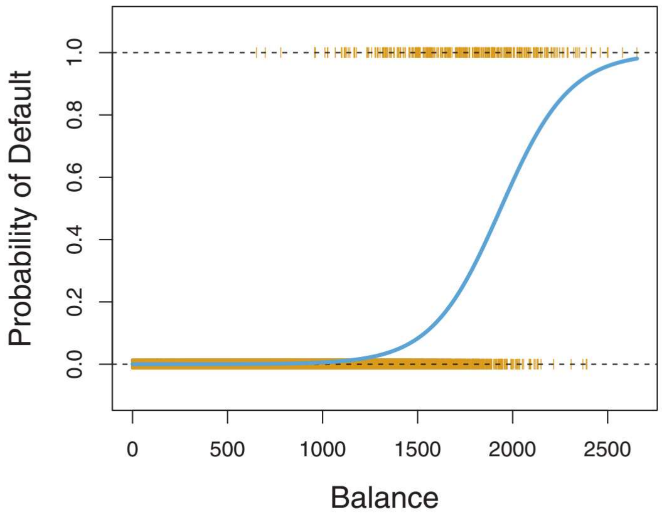
-> the logistic regression model has a logit (or a log-odds) that is linear in X.



Logistic function

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

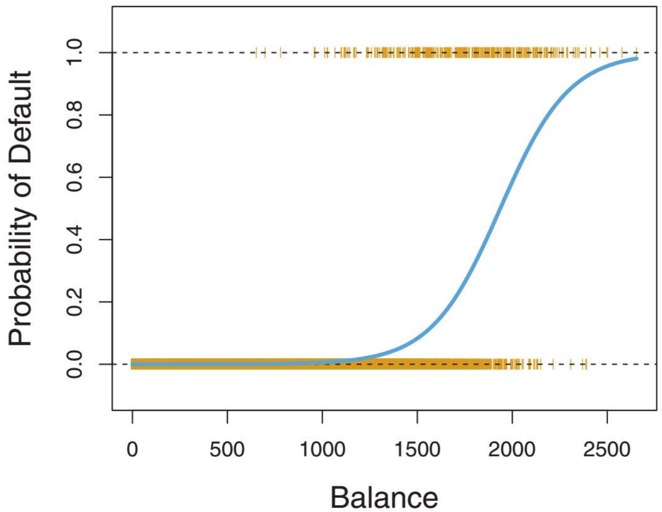
- **What can we say about the average change in Y resulting from a one-unit increase in X?**
    - increasing X by one unit changes the log-odds by β_1
        - i.e., multiplies the odds by e^{β_1}



Logistic function

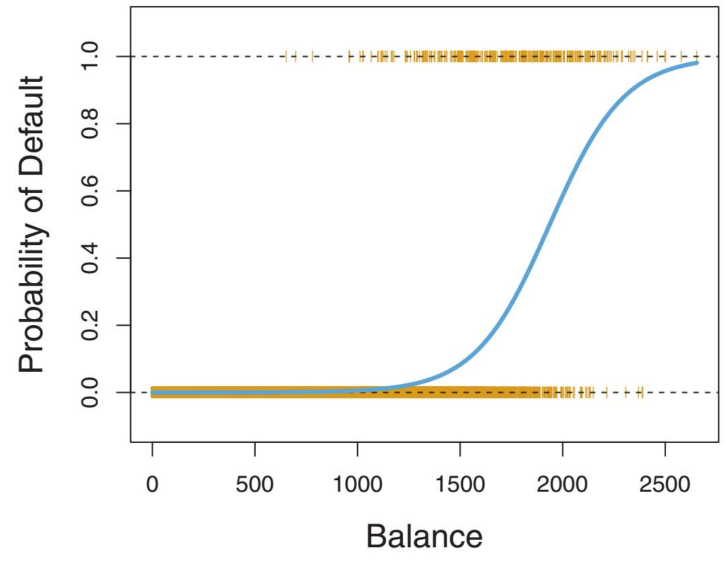
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- **What can we say about the average change in Y resulting from a one-unit increase in X?**



Logistic function

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



- Unlike linear regression, the **relationship between $p(X)$ and X is not a straight line.**

- **$\boldsymbol{\beta_1}$ does not correspond to a unit increase in X!**
  - Variation actually **depends on the current value of X**.

Regardless of the value of X,

if $\beta_1 > 0$, increase in X -> increase in $p(X)$

if $\beta_1 < 0$, increase in X -> decrease in $p(X)$

Estimating the coefficients

- Maximum likelihood method is preferred, due to its statistical properties.

- The estimated values of β_0 and β_1 **need to bring the predicted probability $\hat{\boldsymbol{p}}(\boldsymbol{x_i})$ as close as possible to the actual class of $\boldsymbol{x_i}$.**

  - e.g., in the "Default" data example, $\hat{\beta}_0$ and $\hat{\beta}_1$ should give:
    - a value close to 1 for the persons who defaulted
    - a value close to 0 for the ones who did not

Estimating the coefficients – likelihood function

- The likelihood function is used to estimate the coefficients.

  - It is of the form:

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

    - To be maximized

Estimating the coefficients – likelihood function

- To be maximized

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

	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

Increase in balance -> increase in probability of default.

A unit increase in balance -> increase in the log odds of default by 0.0055 units.

Large absolute values of z-statistic are evidence against the null hypothesis: $H_0: \beta_1 = 0$, i.e., **if large, then there is a relationship between predictor and response**.

Predictions

	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

- What is the predicted default probability for a person with balance of $1000?

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

Predictions

	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

- What is the predicted default probability for a person with balance of $1000?

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

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

for a balance of $2000?

Predictions

	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

- What is the predicted default probability for a person with balance of $1000?

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

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

for a balance of $2000?

$$0.586$$

Predictions

	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

- Logistic regression models can easily accommodate qualitative predictors, e.g., student in this dataset -> 0 or 1.

$$\widehat{\Pr}(\texttt{default}=\texttt{Yes}|\texttt{student}=\texttt{Yes}) =$$

$$\widehat{\Pr}(\texttt{default}=\texttt{Yes}|\texttt{student}=\texttt{No}) =$$

Predictions

	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

- Logistic regression models can easily accommodate qualitative predictors, e.g., student in this dataset -> 0 or 1.

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

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

What can we infer?

Multiple logistic regression

- Suppose we want to predict a binary response using multiple predictors.

- The simple logistic regression model can be generalized into:

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

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

- The **maximum likelihood** method can also be applied to estimate the coefficients.

Multiple logistic regression

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

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

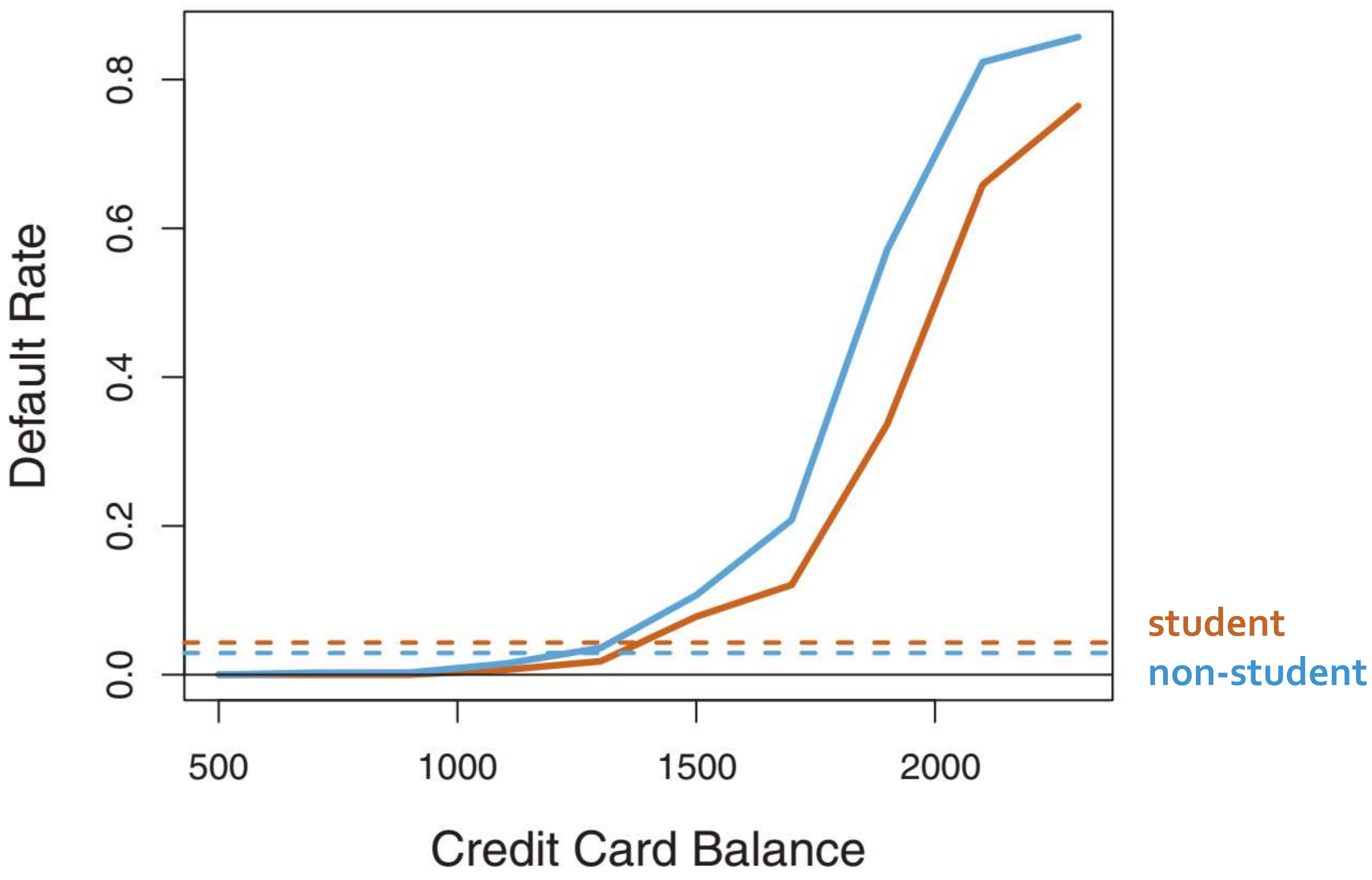
	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

students are less likely to default than non-students

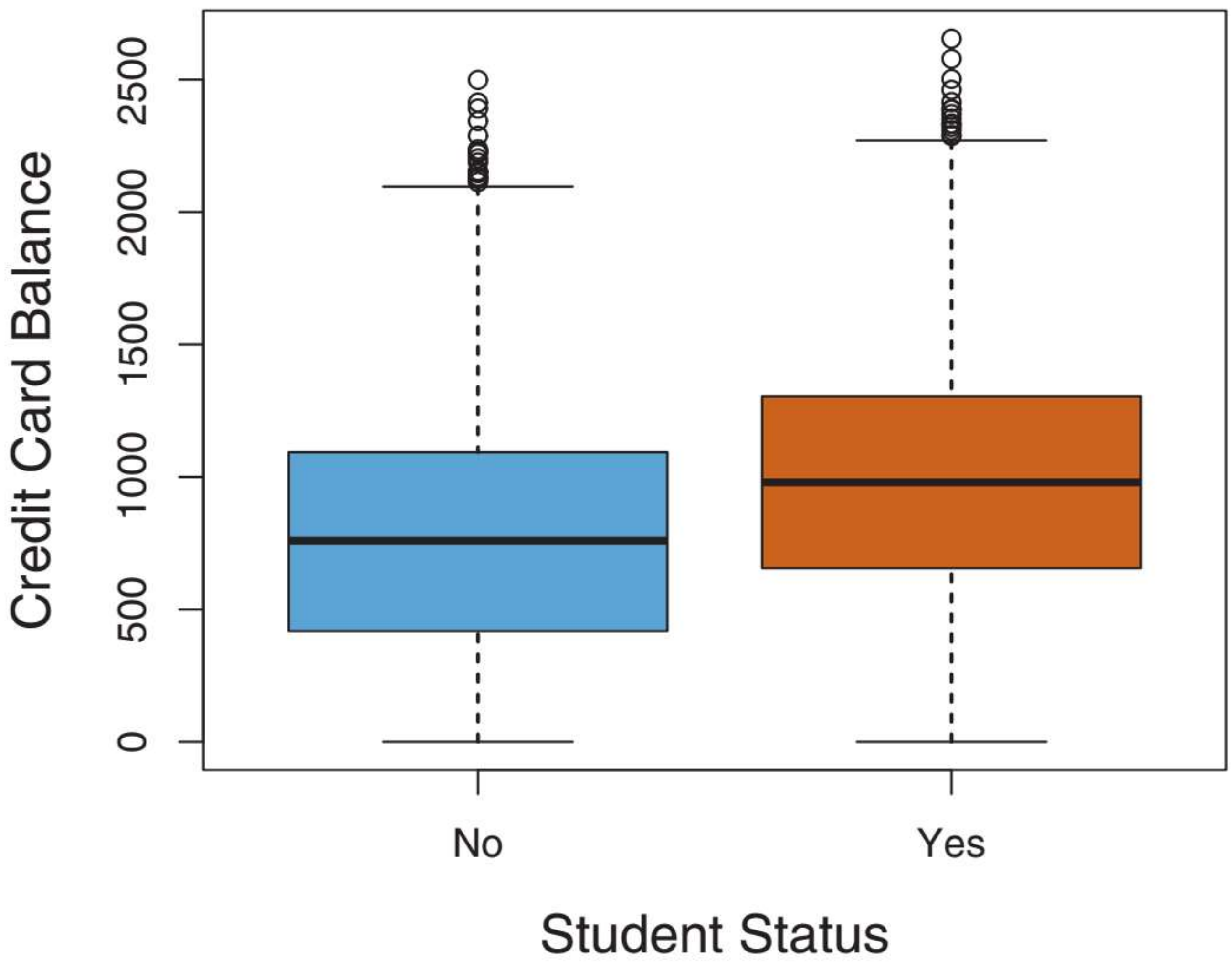
	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

students are more likely to default than non-students

Multiple logistic regression



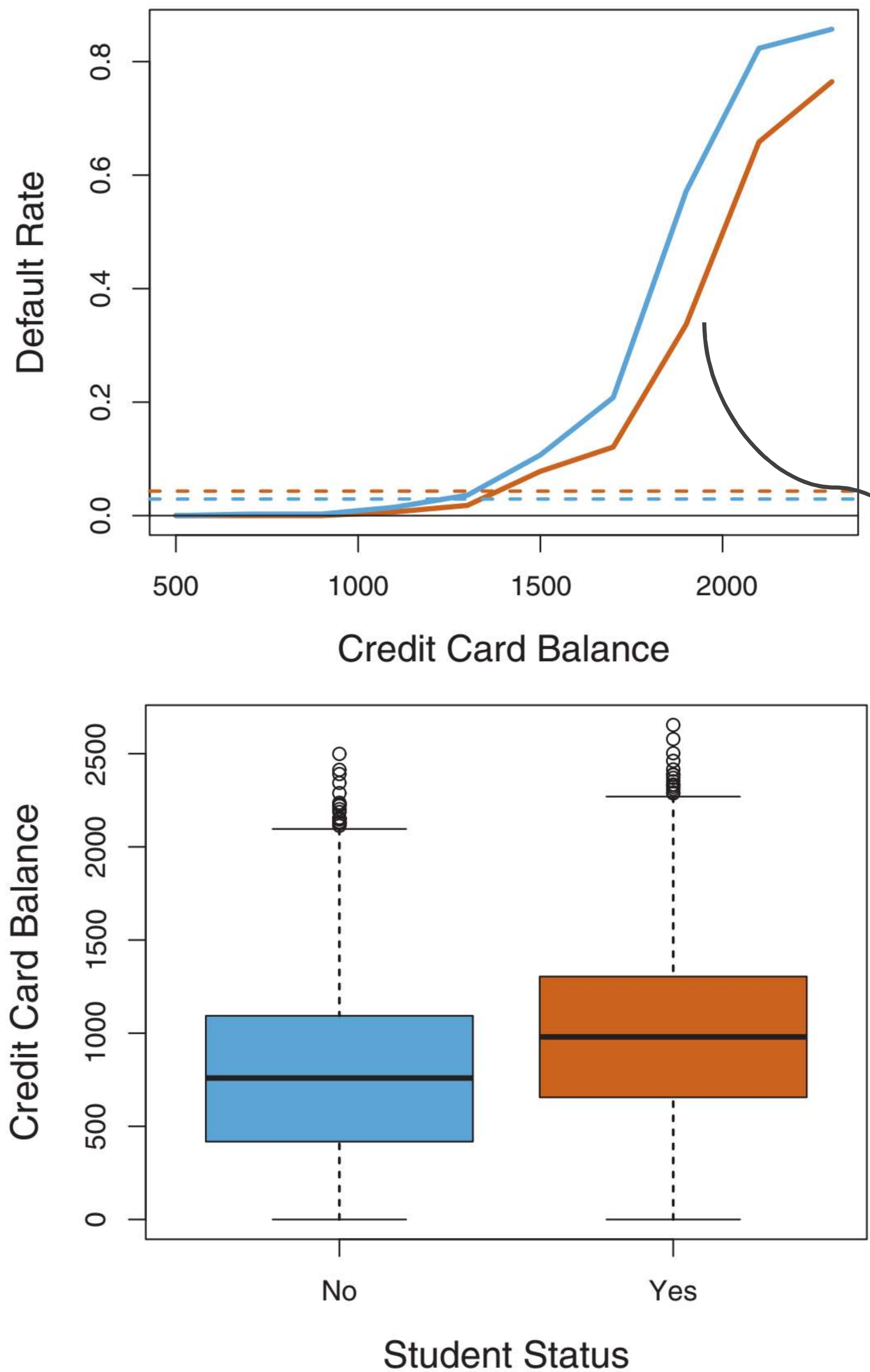
Multiple logistic regression



There is a correlation between student and balance.

-> students tend to have higher balance.

higher balance -> higher default probability!



for the same balance, a student tends to have a lower default probability.

BUT

students tend to have higher balance.

⬇

students tend to have a higher default rate than non-students due to their higher balance.

"Cofounding"

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

	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

What is the predicted default probability for a student with balance of $1500 and income of $40,000?

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

	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

What is the predicted default probability for a student with balance of $1500 and income of $40,000?

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058$$

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

	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

What is the predicted default probability for a non-student with balance of \$1500 and income of \$40,000?

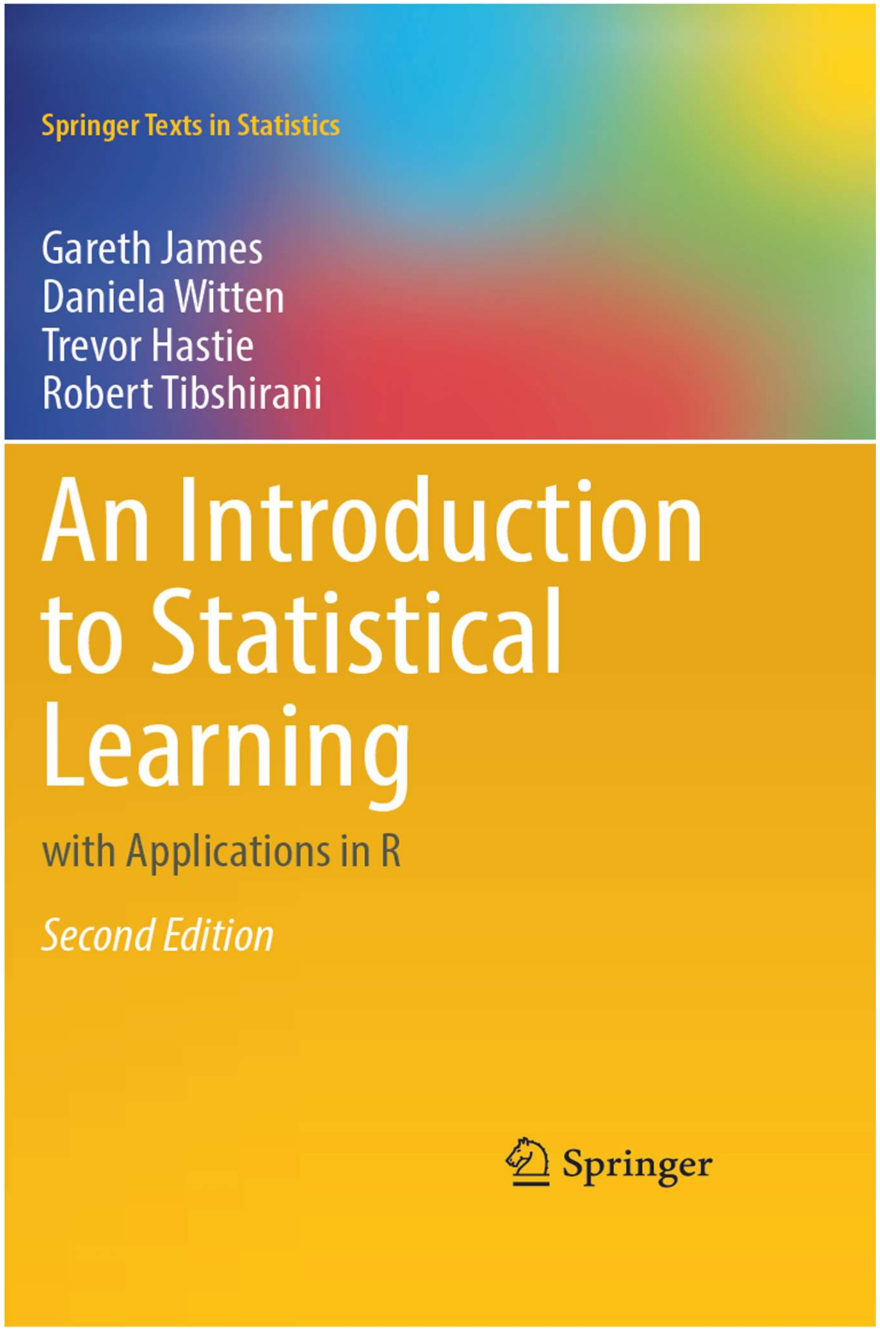
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	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

What is the predicted default probability for a non-student with balance of $1500 and income of $40,000?

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105$$

Reference


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer