LECTURE 06:
CLASSIFICATION PT. 2

September 27, 2017
SDS 293: Machine Learning

Announcements 1/2

- Assignment 1
  - Due **tonight by 11:59pm**
  - Solution set will be posted tomorrow

- **Reminder**: lab posts are due within 24 hours of the lab (i.e. by 10:20am the following day)

Announcements 2/2

- Dr. Kelly Mack visiting Smith **this Friday**
  - VP for Undergraduate STEM Education at the Association of American Colleges and Universities (AAC&U)
  - Former Sr. Program Director for the National Science Foundation (NSF) ADVANCE Program

- Interested in meeting her?
  - Still a few seats available at her morning chat with students
  - Talk to Jordan ASAP to claim one

Outline

✓Motivation

✓Bayes classifier

✓K-nearest neighbors

- Logistic regression
  - Logistic model
  - Estimating coefficients with maximum likelihood
  - Multivariate logistic regression
  - Multiclass logistic regression
  - Limitations
- Linear discriminant analysis (LDA)
  - Bayes' theorem
  - LDA on one predictor
  - LDA on multiple predictors
- Comparing Classification Methods

Flashback: LR vs. KNN for regression

Linear Regression

- Parametric
  - We assume an underlying functional form for $f(X)$
- Pros:
  - Coefficients have simple interpretations
  - Easy to do significance testing
- Cons:
  - Wrong about the functional form → poor performance

K-Nearest Neighbors

- Non-parametric
  - No explicit assumptions about
  - ge about the underlying form
  - More flexible approach
- Cons:
  - Can accidentally “mask” the underlying function

Could a parametric approach work for **classification**?

Example: `default` dataset

	default	student	balance	income
1	No	No	$729.52	$44,361.63
2	No	Yes	$817.18	$12,106.14
⋮	⋮	⋮	⋮	⋮
1503	Yes	Yes	$2232.88	$11770.23
⋮	⋮	⋮	⋮	⋮

Want to estimate:

$$\Pr(\texttt{default} = \texttt{Yes} \mid \texttt{balance})$$

$$\rightarrow p(\mathrm{X})$$

Example: `default` dataset

- How should we model:

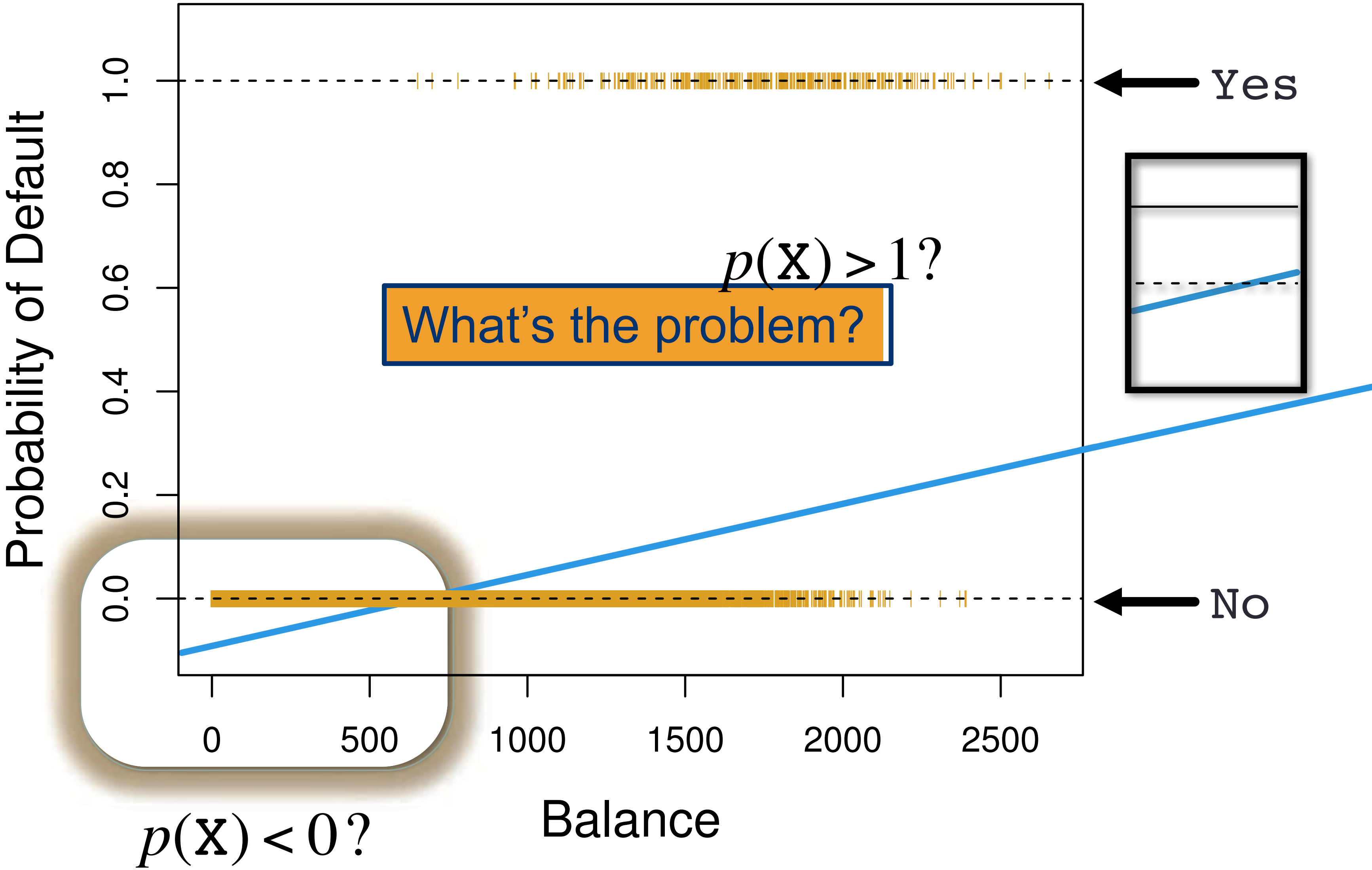
$$p(\texttt{X}) \approx f(\texttt{balance})$$

- We could try a linear approach:

$$p(\texttt{X}) \approx \boxed{\beta_0 + \beta_1}(\texttt{balance})$$

estimate with
linear regression!

Example: `default` dataset



$p(\mathbf{X}) > 1\,?$

What’s the problem?

$p(\mathbf{X}) < 0\,?$

The problem with linear models

- True probability of `default` is bounded by `[0,1]`
- A linear model can always (in theory) predict arbitrarily large/small values
- **Question:** What do we need to fix it?

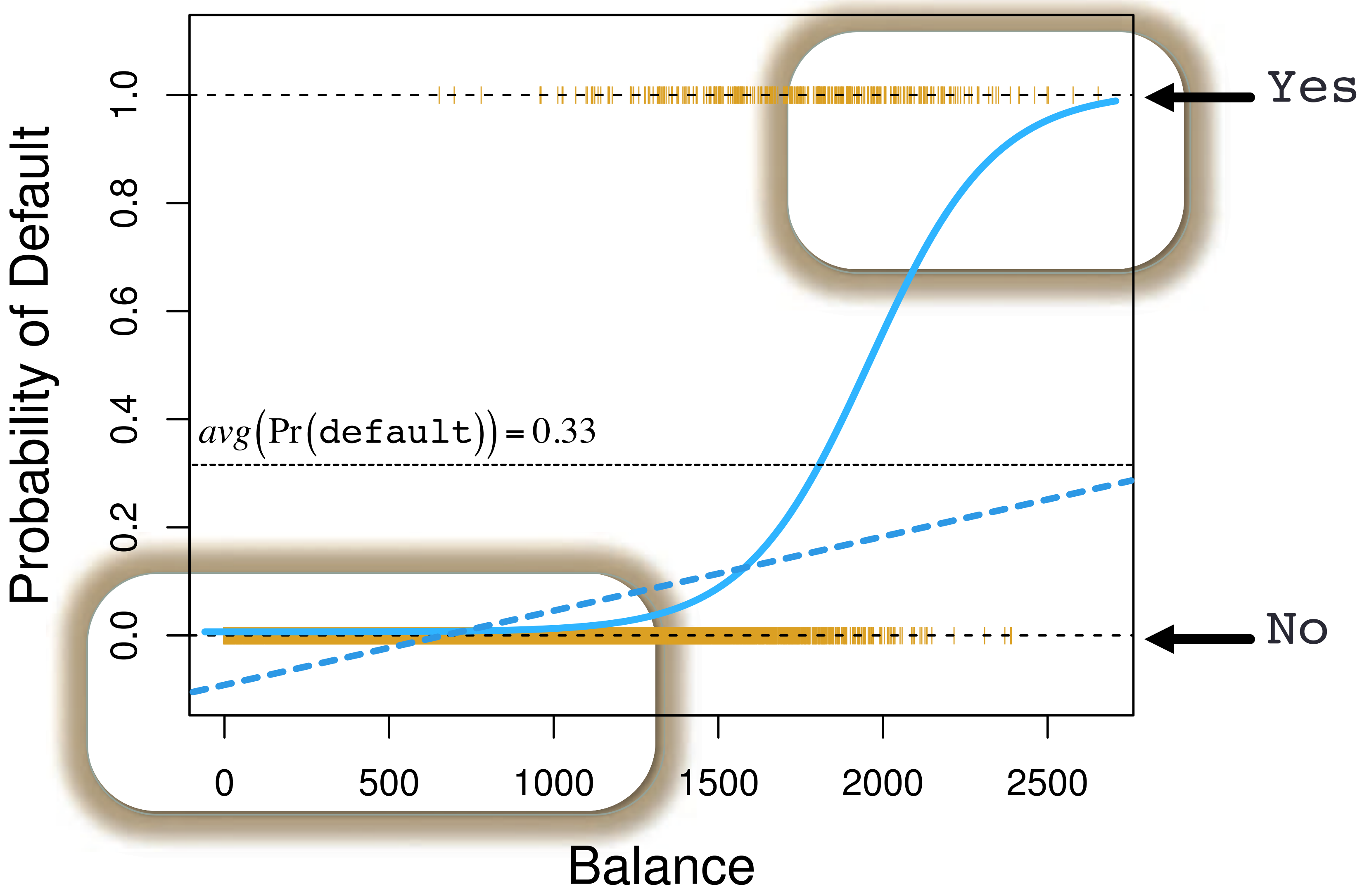
Logistic regression

- **Answer:** a function that gives outputs in `[0,1]` for ALL possible predictor values
- Lots of possible functions meet this criteria (like what?)
- In logistic regression, we use a **logistic function**:

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

*e is the natural logarithm base

Example: `default` dataset



Wait… why a logistic function?

$$\begin{tikzcd} p(\mathrm{X}) = \dfrac{e^{\beta_0+\beta_1 \mathrm{X}}}{1+e^{\beta_0+\beta_1 \mathrm{X}}} \arrow[d, rightsquigarrow] \\ ? = \beta_0 + \beta_1 \mathrm{X} \end{tikzcd}$$

Transforming the logistic model

p(happens)

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

“odds”

p(doesn’t)

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

“log odds”
a.k.a.
logit

$$\boxed{\beta_0 + \beta_1 \mathbf{X}}$$

Flashback: estimating coefficients

- In linear regression, we used **least squares** to estimate our coefficients:

$$\min\left(\sum_{i=0}^{n}(y_i - \hat{y}_i)^2\right)$$

- What's the intuition here?

Estimating coefficients with max likelihood

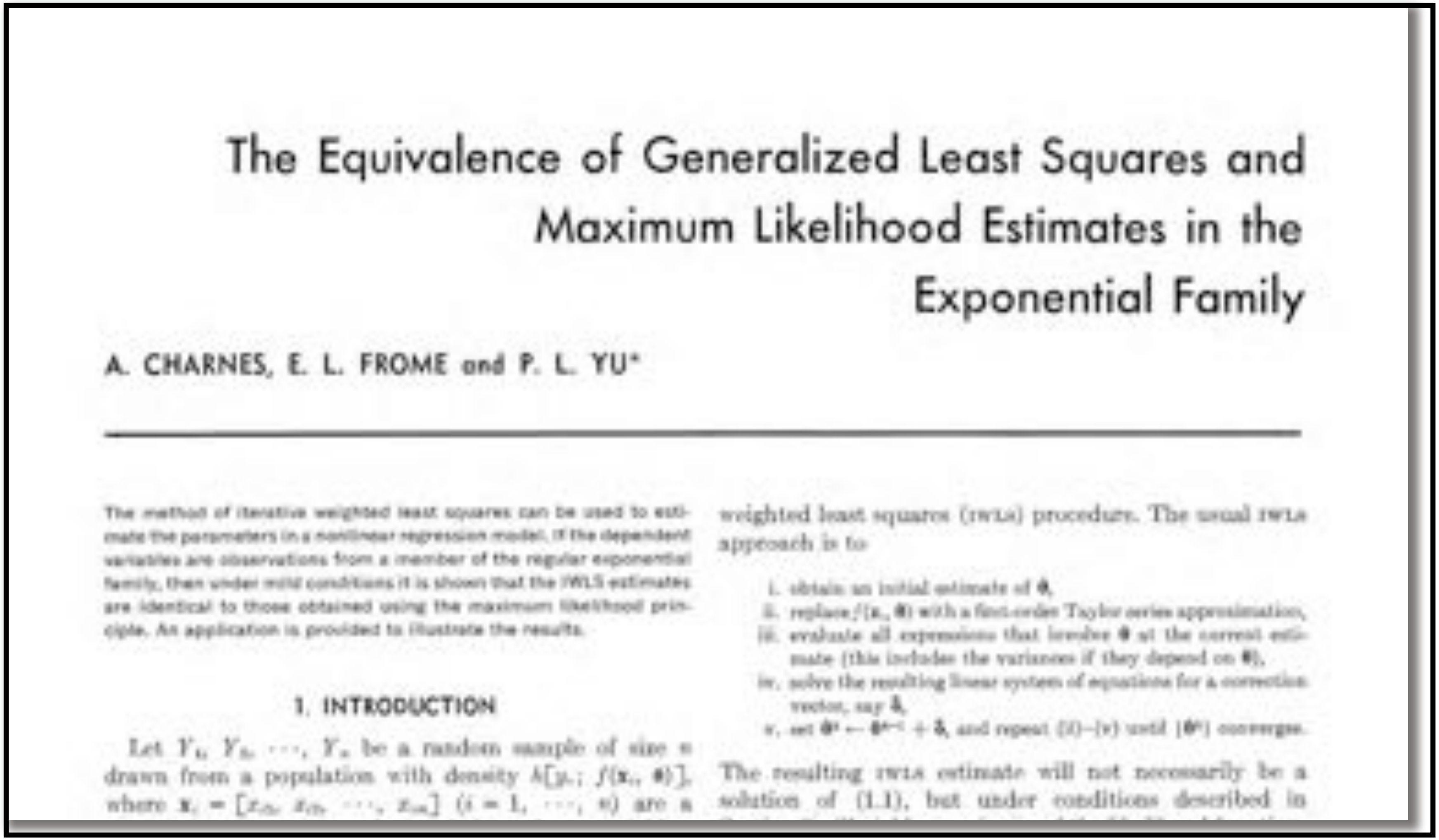
- In logistic regression, we want coefficients such that yield:
  - values close to 1 (high probability) for observations **in the class**
  - values close to 0 (low probability) for observations **not in the class**

- Can formalize this intuition mathematically using a **likelihood** function:

$$\boxed{\prod_{i:y_i=1} p(\mathrm{x}_i)} \times \boxed{\prod_{j:y_j=1} \left(1 - p\left(\mathrm{x}_j\right)\right)}$$

- **Goal**: coefficients that **maximize** this function

Fun fact



The Equivalence of Generalized Least Squares and Maximum Likelihood Estimates in the Exponential Family

A. CHARNES, E. L. FROME and P. L. YU*

The method of iterative weighted least squares can be used to estimate the parameters in a nonlinear regression model. If the dependent variables are observations from a member of the regular exponential family, then under mild conditions it is shown that the IWLS estimates are identical to those obtained using the maximum likelihood principle. An application is provided to illustrate the results.

1. INTRODUCTION

Let $Y_1, Y_2, \cdots, Y_n$ be a random sample of size n drawn from a population with density $h[y_i; f(\mathbf{x}_i, \boldsymbol{\theta})]$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{im}]$ $(i = 1, \cdots, n)$ are a

weighted least squares (IWLS) procedure. The usual IWLS approach is to

i. obtain an initial estimate of $\boldsymbol{\theta}$,
ii. replace $f(\mathbf{x}_i, \boldsymbol{\theta})$ with a first-order Taylor series approximation,
iii. evaluate all expressions that involve $\boldsymbol{\theta}$ at the current estimate (this includes the variances if they depend on $\boldsymbol{\theta}$),
iv. solve the resulting linear system of equations for a correction vector, say $\boldsymbol{\delta}$,
v. set $\boldsymbol{\theta}^k \leftarrow \boldsymbol{\theta}^{k-1} + \boldsymbol{\delta}$, and repeat (ii)–(v) until $\{\boldsymbol{\theta}^k\}$ converges.

The resulting IWLS estimate will not necessarily be a solution of (1.1), but under conditions described in

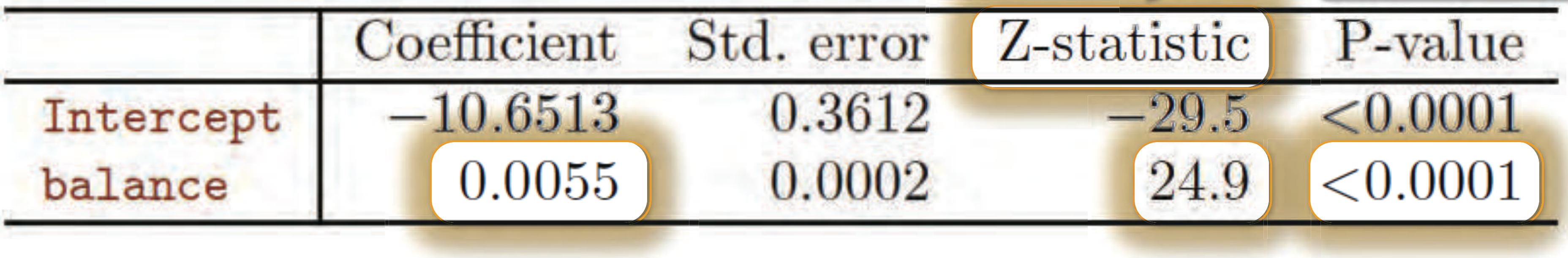
Charnes, Abraham, E. L. Frome, and Po-Lung Yu. "The equivalence of generalized least squares and maximum likelihood estimates in the exponential family." *Journal of the American Statistical Association* 71.353 (1976): 169-171.

Back to our `default` example

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

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

$$\frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)}$$



	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

Making predictions

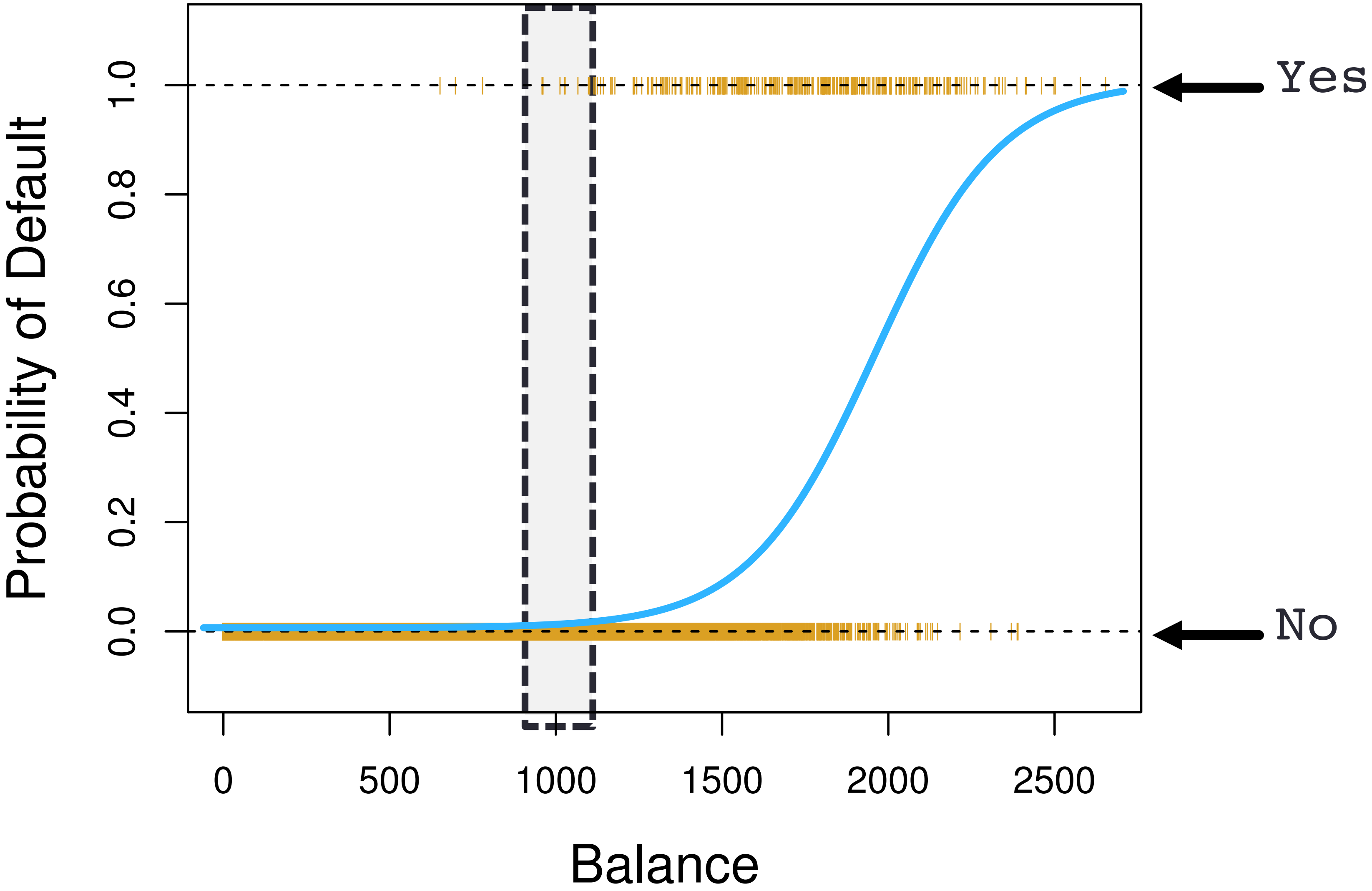
$$\hat{p}(\texttt{default}|\texttt{balance}) = \frac{e^{-10.6513+0.0055(\texttt{balance})}}{1+e^{-10.6513+0.0055(\texttt{balance})}}$$

- Plugging in some sample values:

$$\hat{p}(\texttt{default}|\texttt{\$1000}) = 0.00576 < 1\%$$

$$\hat{p}(\texttt{default}|\texttt{\$2000}) = 0.586 = 58.6\%$$

Example: `default` dataset



Example: `default` dataset

	`default`	`student`	`balance`	`income`
1	No	No	$729.52	$44,361.63
2	No	Yes	$817.18	$12,106.14
⋮	⋮	⋮	⋮	⋮
1503	Yes	Yes	$2232.88	$11770.23
⋮	⋮	⋮	⋮	⋮

How do we handle **qualitative** predictors?

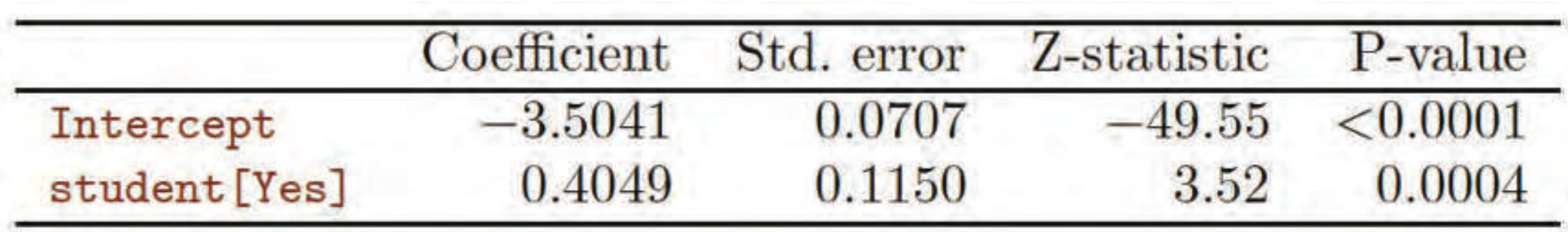
Flashback: qualitative predictors

1
1
0
1
0
1
1
0
1
1

```
({student:“yes”}, {student:“yes”}, {student:“no”},…)
({student[yes]:1}, {student[yes]:1}, {student[yes]:0},…)
```

Qualitative predictors: `default`

$$p(\mathrm{X}) = \frac{e^{\beta_0+\beta_1(\texttt{student[Yes]})}}{1+e^{\beta_0+\beta_1(\texttt{student[Yes]})}}$$



	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

$$p(\texttt{default|student[Yes]}) = 0.0431$$

$$p(\texttt{default|student[No]}) = 0.0291$$

Multivariate logistic regression

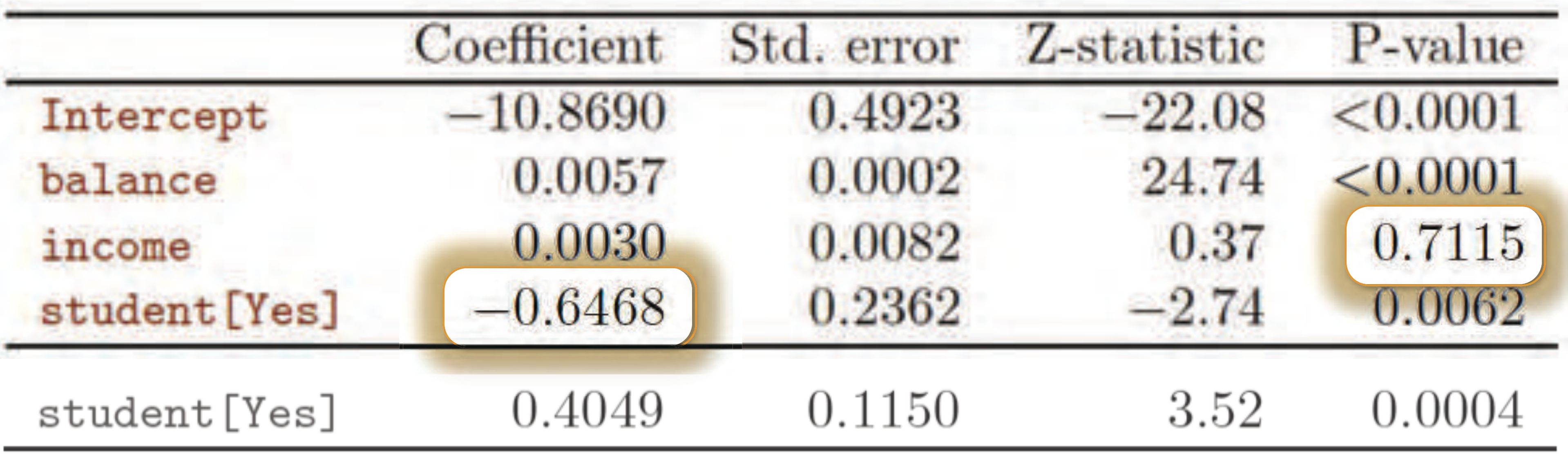
$$log\left(\frac{p(\mathrm{X})}{1-p(\mathrm{X})}\right)=\beta_0+\beta_1(\mathrm{x})$$

$$\downarrow$$

$$log\left(\frac{p(\mathrm{X})}{1-p(\mathrm{X})}\right)=\beta_0+\beta_1(\mathrm{x}_1)+\cdots+\beta_k(\mathrm{x}_k)$$

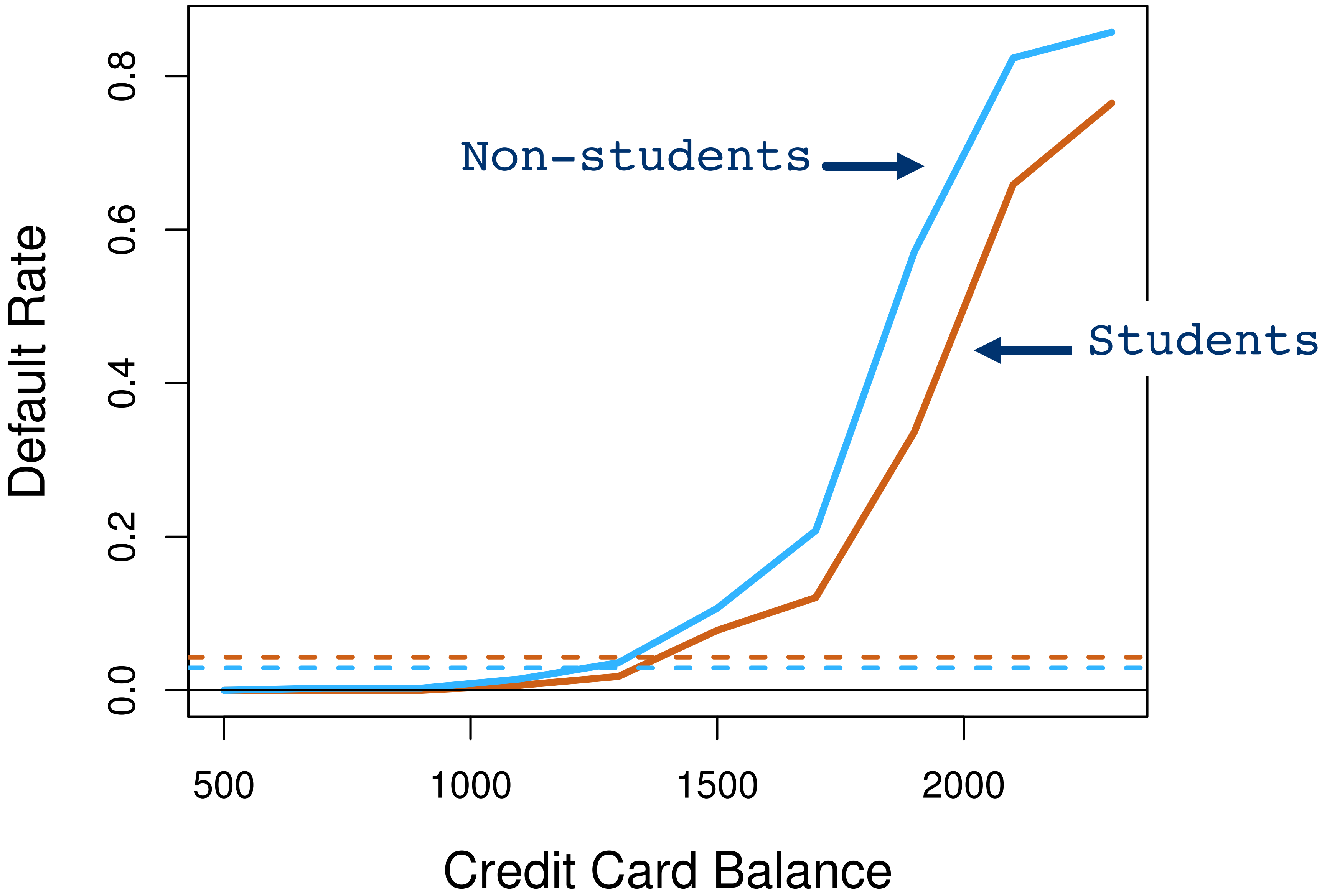
Multivariate logistic regression: `default`

$$log\left(\frac{p(\mathrm{X})}{1-p(\mathrm{X})}\right)=\beta_0+\beta_1(\texttt{balance})$$
$$+\beta_2(\texttt{income})$$
$$+\beta_3(\texttt{student[Yes]})$$

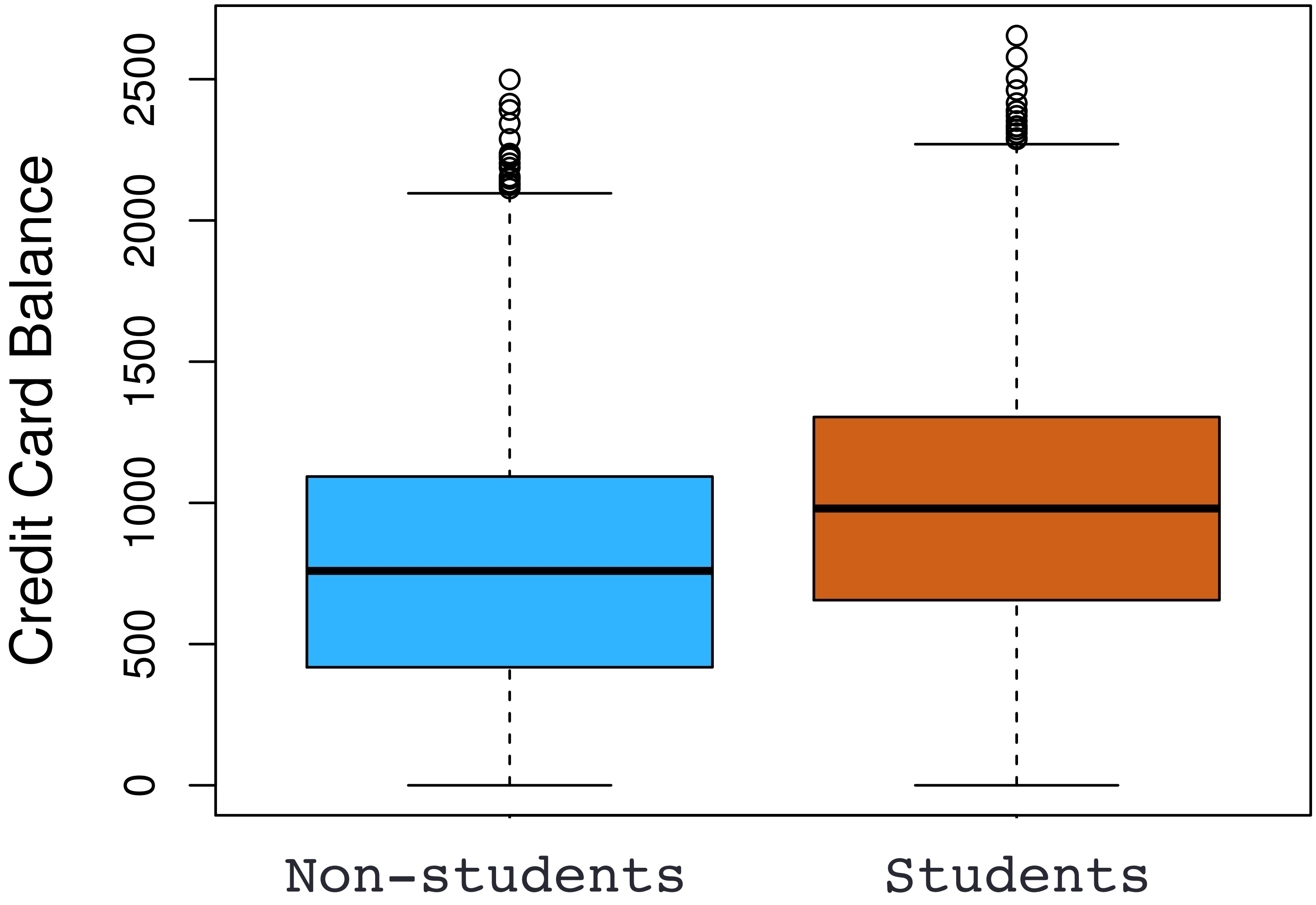


	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062
`student[Yes]`	0.4049	0.1150	3.52	0.0004

What’s going on here?



What's going on here?



What’s going on here?

- Students tend to have a **higher balance** than non-students
- **Higher balances** are associated with **higher default rates** (regardless of student status)
- But if we hold **balance constant**, a student is actually lower risk than a non-student!

This phenomenon is known as
“**confounding**”

Lab: Logistic Regression

- To do today's lab in R: <nothing new>
- To do today's lab in python: `statsmodels`
- Instructions and code:

[course webpage]/labs/lab4-r.html

[course webpage]/labs/lab4-py.html

- Full version can be found beginning on p. 156 of ISLR

Recap: binary classification

👍 logistic regression

What if we have multiple classes?

MARVEL DC DRAGONBALL Z

Multiclass logistic regression

- The approach we discussed today has a straightforward extension to more than two classes
- R will build you multiclass logistic models, no problem
- **In reality:** they're almost never used
- Want to know why? Come to class on Monday ☺

Coming up

- Monday: **linear discriminant analysis**
  - Bayes’ theorem
  - LDA on one predictor
  - LDA on multiple predictors
- A2 posted, due **Oct. 4th by 11:59pm**

Backup: logit derivation

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

$$= e^{\beta_0+\beta_1\mathtt{X}} \times \left(\frac{1}{1+e^{\beta_0+\beta_1\mathtt{X}}}\right)$$

$$= e^{\beta_0+\beta_1\mathtt{X}} \times \left(\frac{1+e^{\beta_0+\beta_1\mathtt{X}} - e^{\beta_0+\beta_1\mathtt{X}}}{1+e^{\beta_0+\beta_1\mathtt{X}}}\right)$$

$$= e^{\beta_0+\beta_1\mathtt{X}} \times \left(\frac{1+e^{\beta_0+\beta_1\mathtt{X}}}{1+e^{\beta_0+\beta_1\mathtt{X}}} - \frac{e^{\beta_0+\beta_1\mathtt{X}}}{1+e^{\beta_0+\beta_1\mathtt{X}}}\right)$$

$$= e^{\beta_0+\beta_1\mathtt{X}} \times \left(1 - p(\mathtt{X})\right)$$

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

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