Statistical modeling in logistics

MMS075

Lecture 6a – Logistic regression (cont.)

Plots and ethical aspects

Multivariate case, confounding

Multinomial logistic regression

Acknowledgement: Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Outline

- Assignments and exam
  - Assignment 1 – intended solution
  - Assignment 2 information
  - Exam guidelines
- Logistic regression (cont.)
  - Plots and ethical aspects
  - Multivariate vs single variable model
  - Confounding
- Multinomial (i.e. multiclass) logistic regression
- Feed-forward

Recommended resources

Reading in ISL: Sections 4.1-4.3 for theory, 4.6.1, 4.6.2, end of 4.6.6 for R codes

Online resources for multiclass logistic regression:

https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

The videos from the Statistical Learning course are available at this link. Relevant videos for the new material today:

- **Multivariate Logistic Regression and Confounding** (9:53)
- **Case-Control Sampling and Multiclass Logistic Regression** (7:28)
- **Lab: Logistic Regression** (10:14)

Assignments and exam

Assignment 1 – intended solution

Assignment 2 information

Exam guidelines

Assignment 1 – intended solution

1. Is there a relationship between ice cream consumption and at least one of the variables Income, Price and Temperature?

Consider a multiple linear regression model:

$$\text{Consumption} = \beta_0 + \beta_I \times \text{Income} + \beta_P \times \text{Price} + \beta_T \times \text{Temperature} + \varepsilon$$

Test model significance:

$$H_0 : \beta_I = \beta_P = \beta_T = 0$$

$$H_a : \text{at least one of } \beta_I, \beta_P \text{ and } \beta_T \text{ is not zero}$$

```
Call:
lm(formula = Consumption ~ Income + Price + Temperature)

Residuals:
      Min        1Q    Median        3Q       Max
-0.065302 -0.011873  0.002737  0.015953  0.078986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1973151  0.2702162   0.730  0.47179
Income       0.0033078  0.0011714   2.824  0.00899 **
Price       -1.0444140  0.8343573  -1.252  0.22180
Temperature  0.0034584  0.0004455   7.762  3.1e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03683 on 26 degrees of freedom
Multiple R-squared:  0.719,     Adjusted R-squared:  0.6866
F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

Reject H_0, conclude that at least one variable has a relationship with consumption

Assignment 1 – intended solution (cont.)

2. Do all the predictors help to explain ice cream consumption?

Test variable significance in the same model, for each variable:

$$H_0 : \beta_I = 0$$
$$H_a : \beta_I \neq 0$$

Reject H_0, conclude that income helps to explain consumption

$$H_0 : \beta_P = 0$$
$$H_a : \beta_P \neq 0$$

Cannot reject H_0 → price does not help to explain consumption

$$H_0 : \beta_T = 0$$
$$H_a : \beta_T \neq 0$$

Reject H_0, conclude that temperature helps to explain consumption

```
Call:
lm(formula = Consumption ~ Income + Price + Temperature)

Residuals:
      Min        1Q    Median        3Q       Max
-0.065302 -0.011873  0.002737  0.015953  0.078986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1973151  0.2702162   0.730  0.47179
Income       0.0033078  0.0011714   2.824  0.00899 **
Price       -1.0444140  0.8343573  -1.252  0.22180
Temperature  0.0034584  0.0004455   7.762  3.1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.03683 on 26 degrees of freedom
Multiple R-squared:  0.719,     Adjusted R-squared:  0.68
F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

Assignment 1 – intended solution (cont.)

3. Two separate models, one with only Income, other one with Income & Temperature as predictors. Analyze the role of Income.

Check p-value for t-test for Income in each model:

```
Call:
lm(formula = Consumption ~ Income)

Residuals:
      Min        1Q    Median        3Q       Max
-0.100606 -0.050393 -0.009145  0.034013  0.185840

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.316715   0.168665   1.878   0.0709 .
Income      0.000505   0.001988   0.254   0.8014
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06688 on 28 degrees of freedom
Multiple R-squared:  0.002298,  Adjusted R-squared:  -0.03333
F-statistic: 0.06449 on 1 and 28 DF,  p-value: 0.8014
```

Very high p-value → income in itself does not help to explain consumption

```
Call:
lm(formula = Consumption ~ Income + Temperature)

Residuals:
      Min        1Q    Median        3Q       Max
-0.065420 -0.022458  0.004026  0.015987  0.091905

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.113195   0.108280  -1.045  0.30511
Income       0.003530   0.001170   3.017  0.00551 **
Temperature  0.003543   0.000445   7.963 1.47e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03722 on 27 degrees of freedom
Multiple R-squared:  0.7021,    Adjusted R-squared:  0.68
F-statistic: 31.81 on 2 and 27 DF,  p-value: 7.957e-08
```

Very low p-value → once temperature is included in the model, income can improve the model and help to explain consumption

Assignment 1 – intended solution (cont.)

4. What consumption of ice cream per head do you predict assuming an average family income of 85 dollars and an average temperature of 50 Fahrenheit? What would be your prediction at the same income and an average temperature of 70 Fahrenheit?

- Solution will be provided in a document on Canvas once all re-submissions are complete

- Can be answered either using software or calculation by hand, using software output on previous slide

Assignments and exam

Assignment 1 – intended solution

Assignment 2 information

Exam guidelines

Assignment 2 information

- Assignment 2 has been published in Canvas, deadline: Monday, March 2, 23:59. Make sure to submit a solution before the deadline, even if it is imperfect

- You can use any software you want to get the solution – part 3 is formulated in terms of R, but appropriate discussion of diagnostic plots from other software is also accepted

- Checking lectures & computer labs in weeks 3, 4, 5 should suffice for solving the assignment; feel free to use consultation times if you want to discuss the material

Assignments and exam

Assignment 1 – intended solution

Assignment 2 information

Exam guidelines

Assessment in MMS075

The description of assessment on the Student Portal:

Examination including compulsory elements

One or more project tasks (part A). Written examination (part B). The final grade is determined by the grade on the written exam.

Credit distribution

Module			Sp1	Sp2	Sp3	Sp4	Summer course	No Sp	Examination dates		
0119	Written and oral assignments, part A	5,0c Grading: UG		5,0c							
0219	Examination, part B	2,5c Grading: TH		2,5c					16 Mar 2020 am L,	10 Jun 2020 pm L,	19 Aug 2020 am L

Link between assessment and intended learning outcomes

Learning outcomes (After the course, students should be able to...)

A+E • Demonstrate an understanding of the key concepts and ideas in statistical modeling on larger datasets;

E • Describe suitable statistical methods for using on larger datasets relevant in logistics;

A+E • Choose and use appropriate statistical methods for answering a logistics related problem, and report the findings in a suitable and compelling format;

A+E • Critically evaluate statistical materials and methods and reason about their limitations;

A • Reflect on ethical aspects and considerations when collecting and analyzing larger datasets.

A: assignments, E: exam

Will the exam include ...

- **Computations** by hand or using a calculator?

No long computations.

Short computations (of 1-2 steps) are possible, formulas will be provided.

- **Choosing models** and **interpreting coefficients**?

Yes. You need to understand when the different models can be used and how the model parameters can be interpreted.

- Interpretation of **outputs from R**?

Yes. It is essential to understand summaries and plots provided by statistical software like R and find the relevant information in such outputs.

- Specification or interpretation of **commands from R**?

No. The course is about statistical modelling and not about a specific software.

Logistic regression (cont.)

Plots and ethical aspects

Multivariate vs single variable case

Confounding

Recall: binary response

A response is binary if there are two possible outcomes:

- Broken leg or not broken
- Default on credit card or not default
- An email is spam or not spam

Note: this classification excludes any other health conditions; for example, a person with brain damage and a broken arm but without a fractured leg would be in the ”not broken” category

The outcome of interest is called a **”case”** or a **”success”** (even if it is the worse outcome, like spam email) and the response can be coded as follows:

$$Y = \begin{cases} 1 \text{ if the response is a “case”, i.e. the outcome that we want to analyse} \\ 0 \text{ otherwise} \end{cases}$$

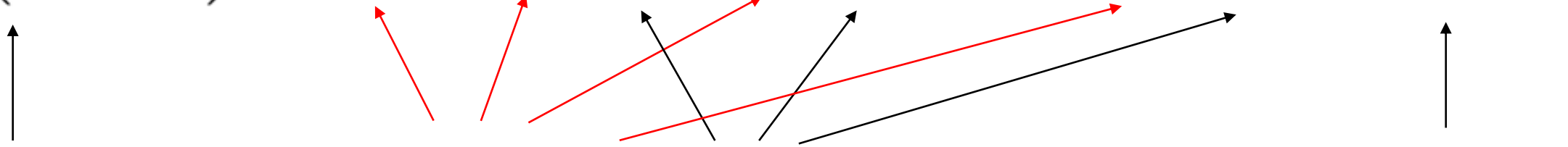
For given values of all predictors: $X = (X_1, X_2, \ldots, X_p)$, we denote **the probability of a case** by $p(X)$, i.e. we define the following conditional probability:

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

Logistic regression model (binomial case)

- Log-odds of a ”case” linearly depends on $p \geq 1$ predictors:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$



Log-odds of a ”case”
Log-odds of ”success”
Log-odds of outcome 1
Same expressions with logit instead of log-odds

Coefficients
Parameters

Predictors
Features
Input variables
Independent variables

Error term

- For non-binary response: multinomial logistic regression (last slides)

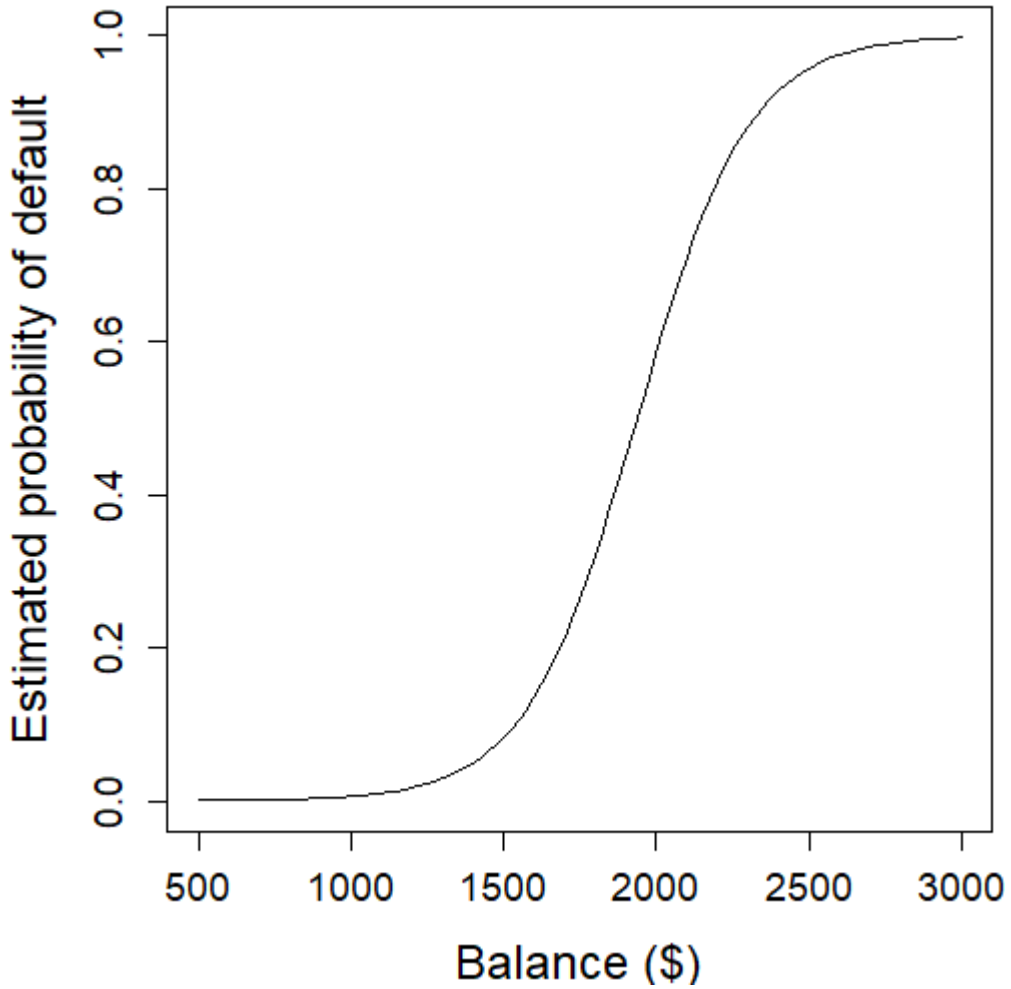
Example: balance as predictor

- We want to analyse the probability of default at given values of the monthly credit card balance

- Software gives coefficient estimates using a method called maximum likelihood, see R output below:

```
Call:
glm(formula = default ~ balance, family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

Coefficients:
             Estimate  Std. Error z value            Pr(>|z|)
(Intercept) -10.6513306  0.3611574  -29.49 <0.0000000000000002 ***
balance       0.0054989  0.0002204   24.95 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5

Number of Fisher Scoring iterations: 8
```

Effect of extra $1 in balance on defaulting



The effect of one-unit increase in the predictor,
i.e. $1 increase in balance, has the following effects:

- Increases log-odds of default by 0.0055
- Multiplies odds of default by $e^{0.0055} = 1.005515$
- Increases probability of default by some amount

This depends on where we are on the x-axis!

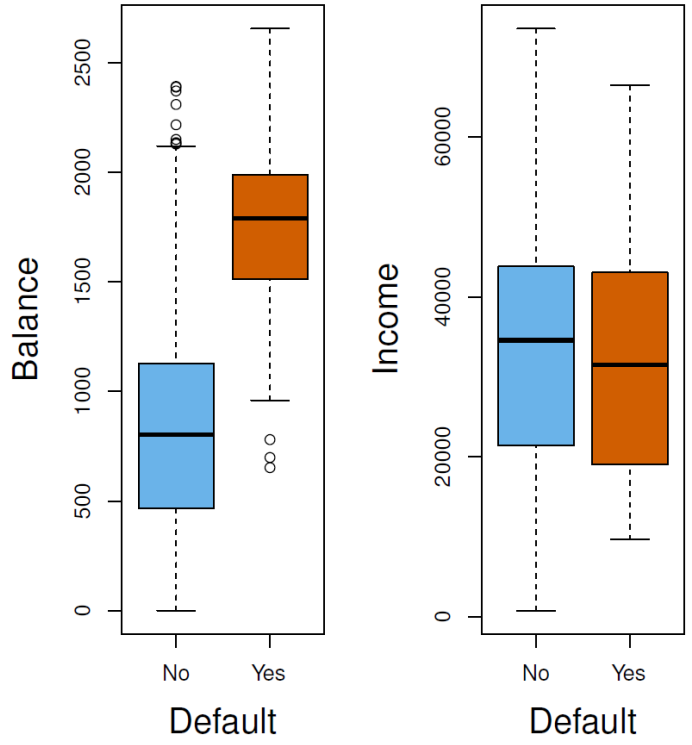
- Under $1500, the $1 extra in balance has little effect
- Between $1500 and $2500, it has larger effect
- Above $2500, it has little effect

An **S-shaped curve** is typical in logistic regression

How can box plots help the analysis?

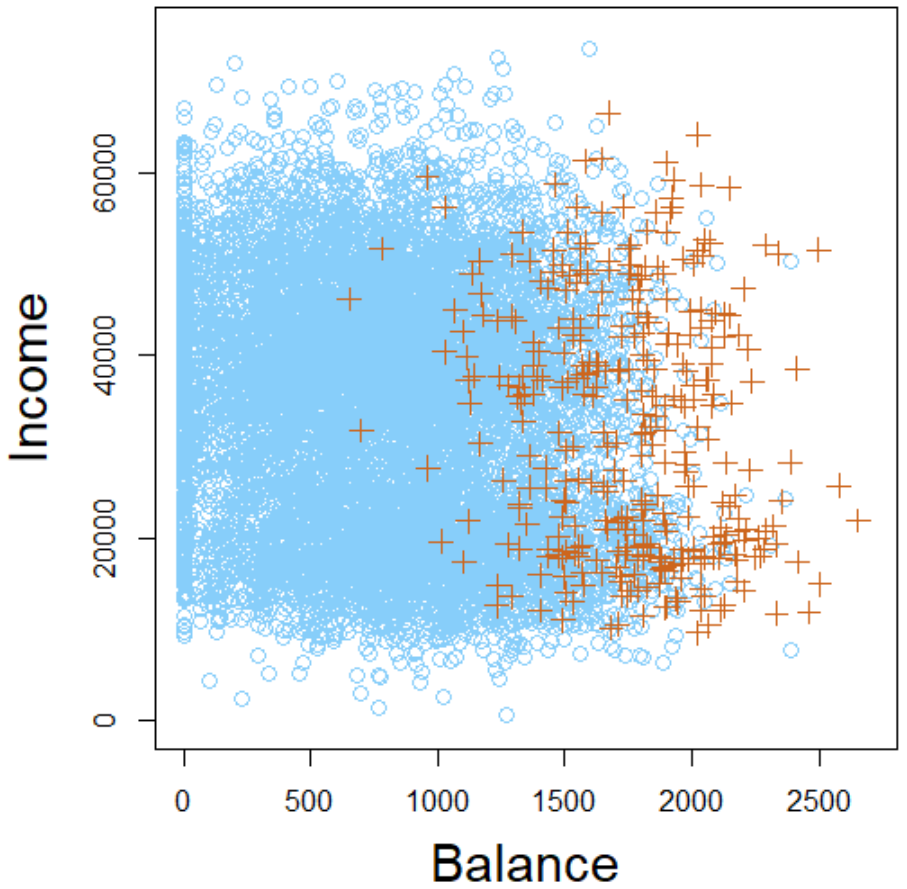
- Box plots can suggest variables for model:
  - Very different balance levels for defaulted vs not defaulted → balance can help to predict default
  - Similar income for defaulted and non-defaulted → income does not seem equally important

- Box plots do not indicate number of defaulted or non-defaulted people → they give very little information about probability of default



How can scatter plots help the analysis?

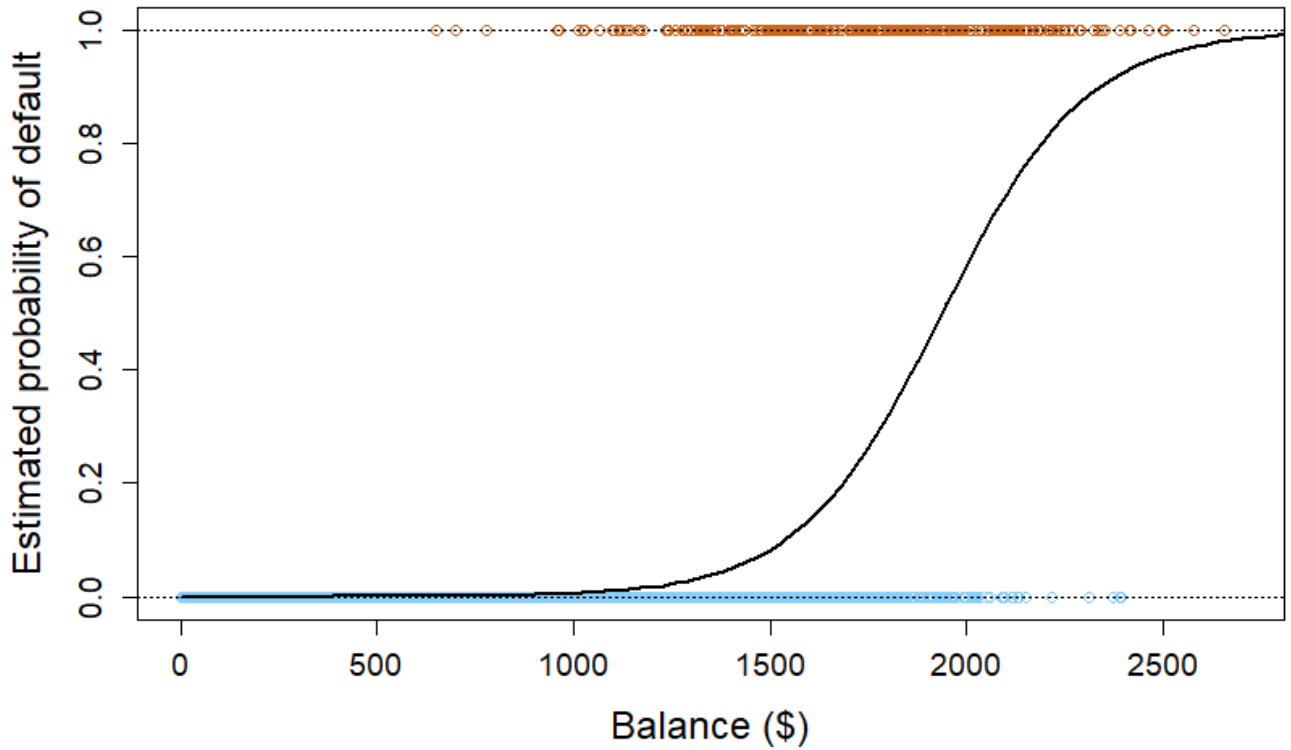
- Plot cases and non-cases for predictor combinations, look for patterns:
  - Clear tendency of ”blue points more to the left, brown points more to the right” → x-axis variable may be a good predictor
  - Less clear pattern for vertical orientation → y-axis variable may be less important
- Relative frequency of default cases around given balance values may be close to estimated probability



Plotting results (single numerical predictor)

- Predictor values on x-axis, probability of a case on y-axis

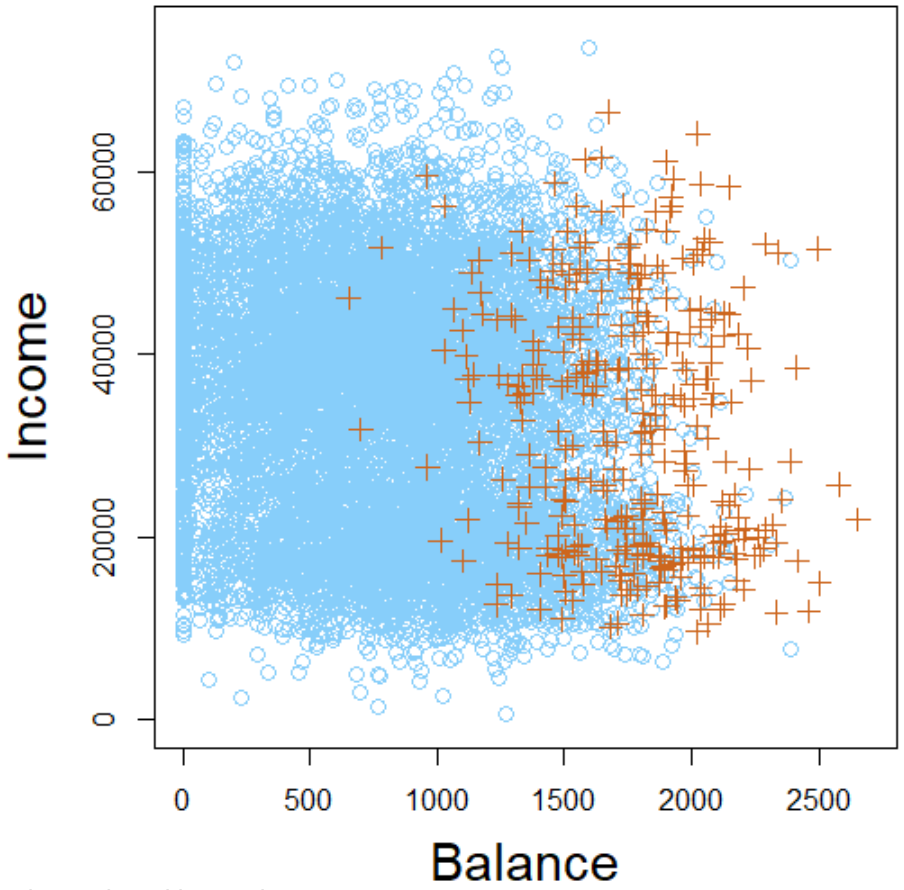
- Scatter plot of re-coded response Y (value 1 for cases and 0 for non-cases) versus predictor is often added to the plot (see brown and blue points)



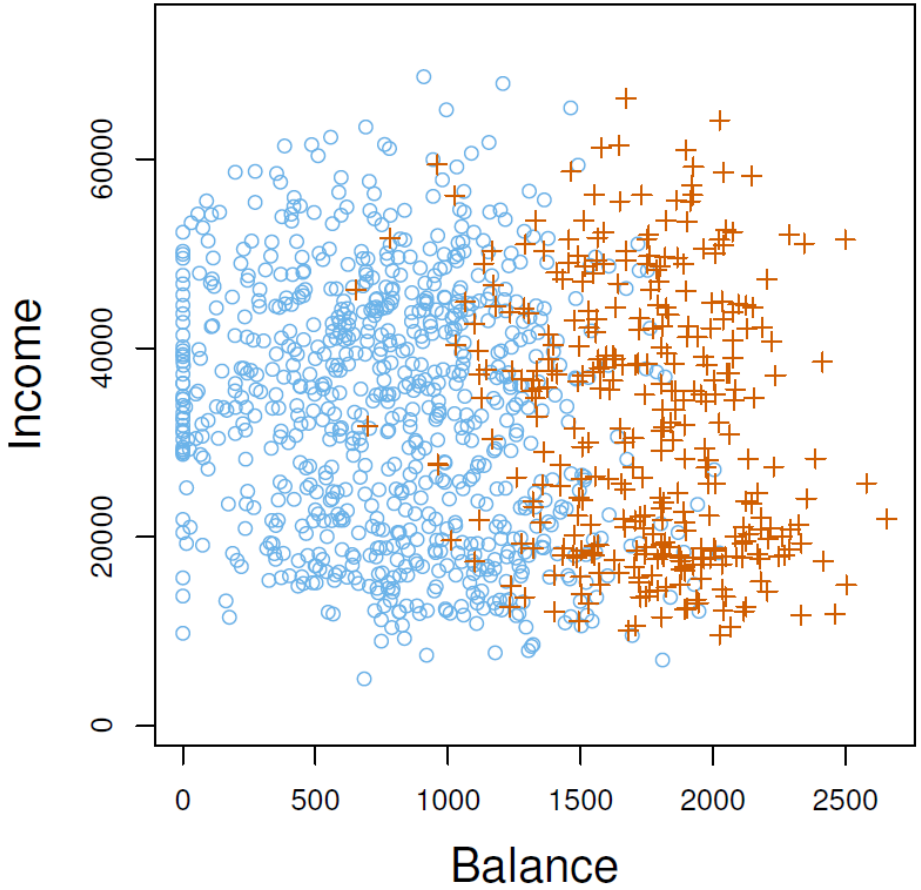
Showing all data vs tidier figure

In Figure 4.1 in ISL, only a fraction of non-default cases is shown:

Scatter plot Income vs Balance in Default dataset in R



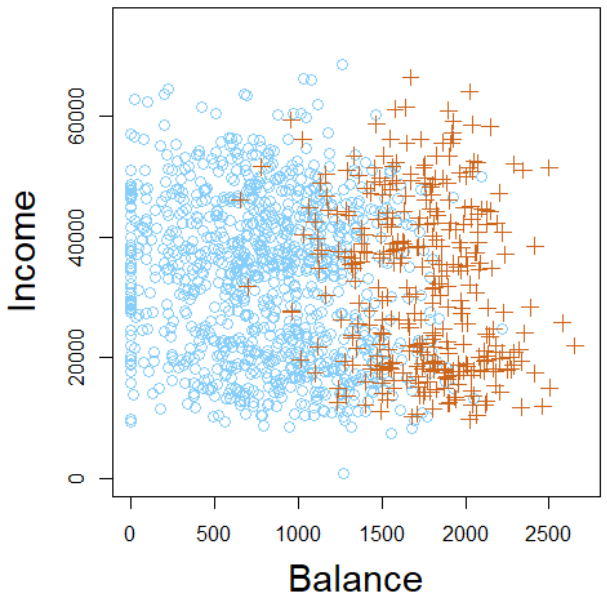
Source: Figure 4.1 in ISL



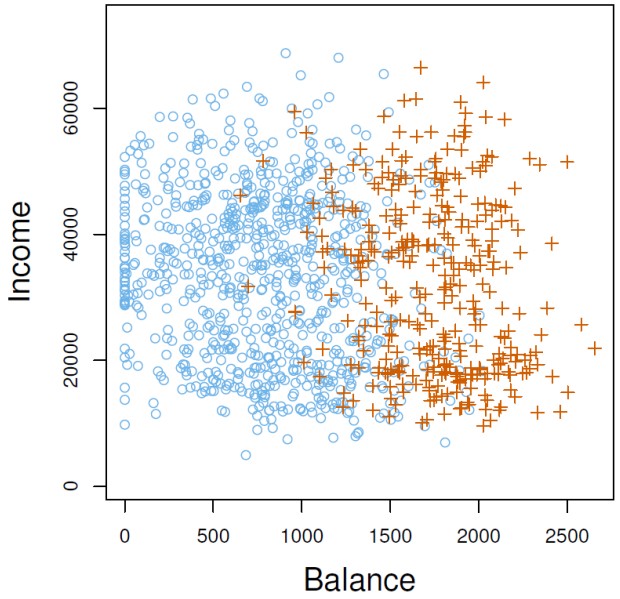
What model would ISL scatter plot give?

- How to get ISL plot from the scatter plot with all data? Removing 90% of non-defaulted individuals gets close:

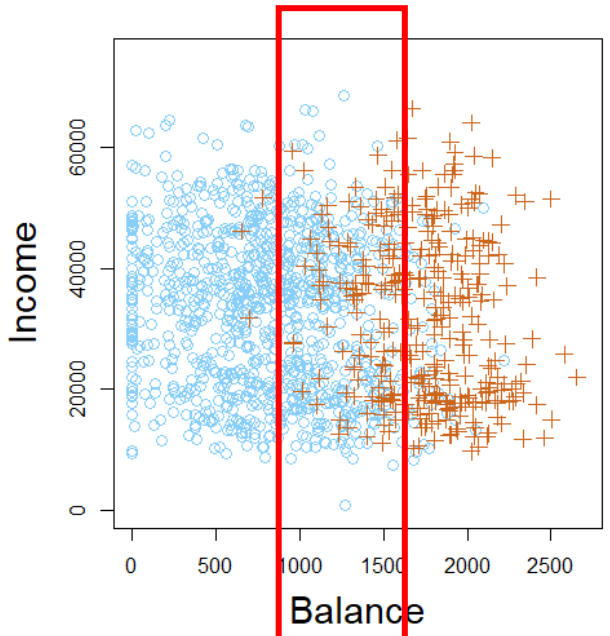
Scatter plot after removing 90% of non-default points

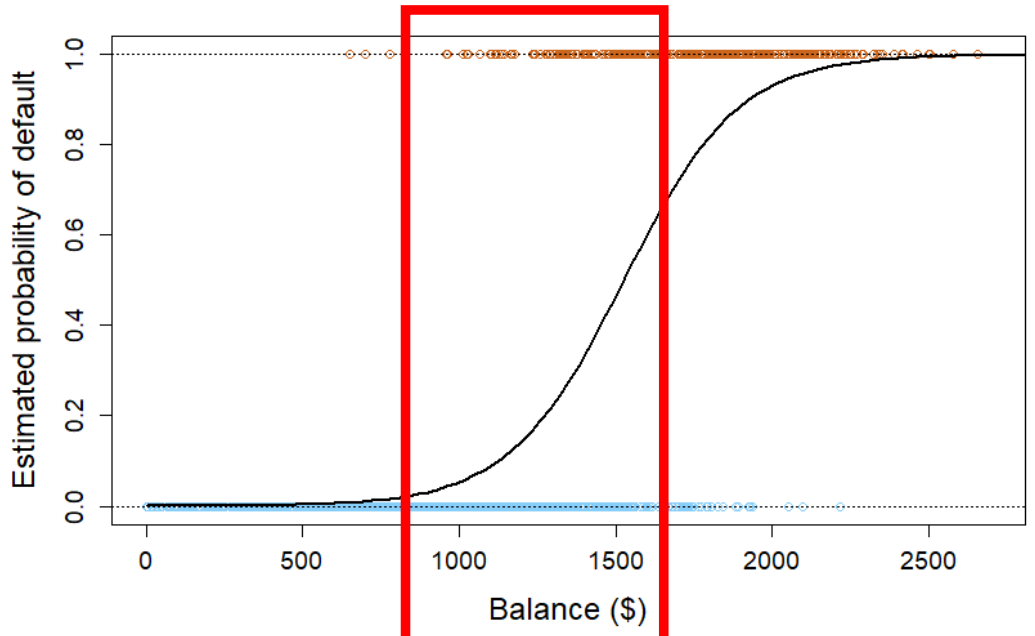


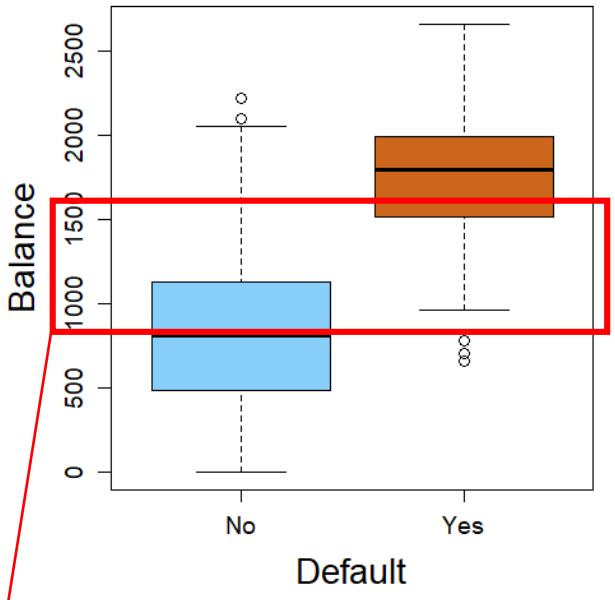
Source: Figure 4.1 in ISL

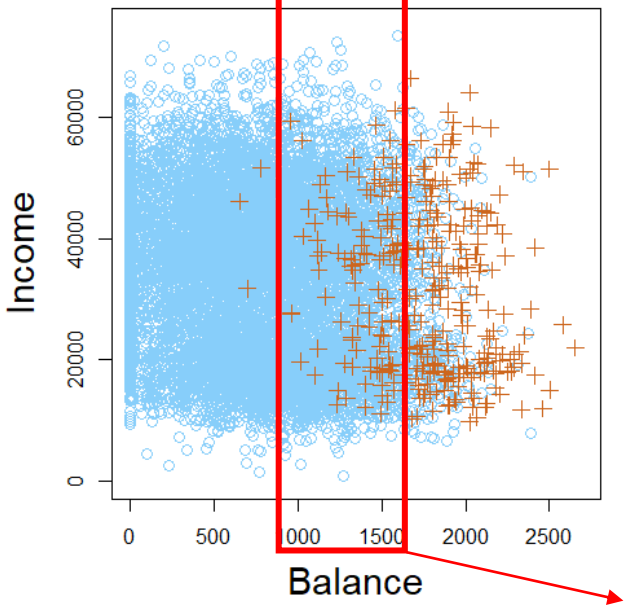


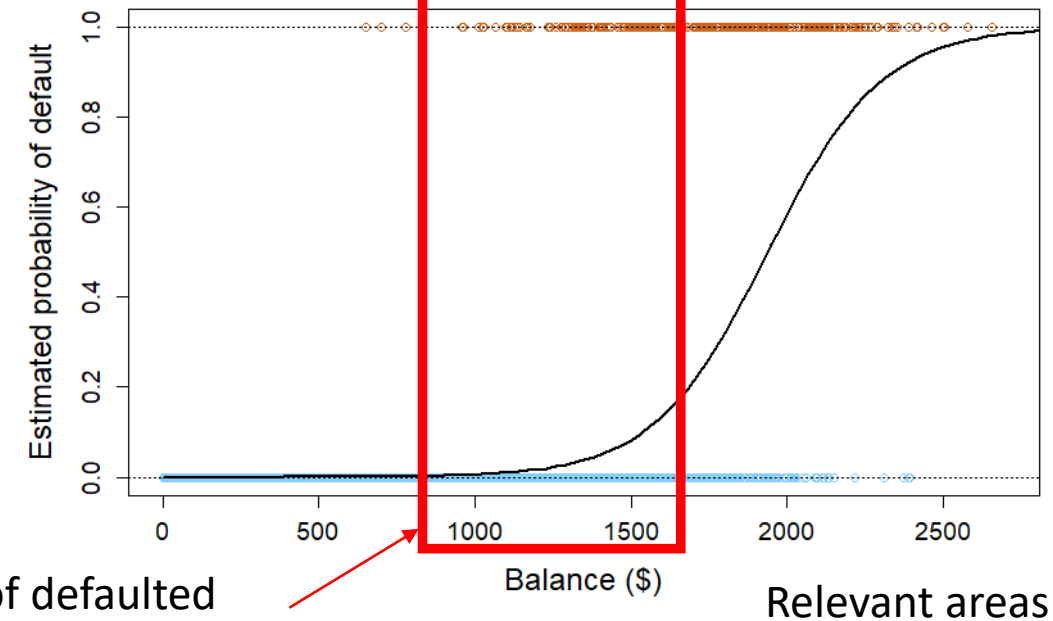
- Next slide: probability of default curves and box plots for both model with removing 90% of non-defaulted & model with all data

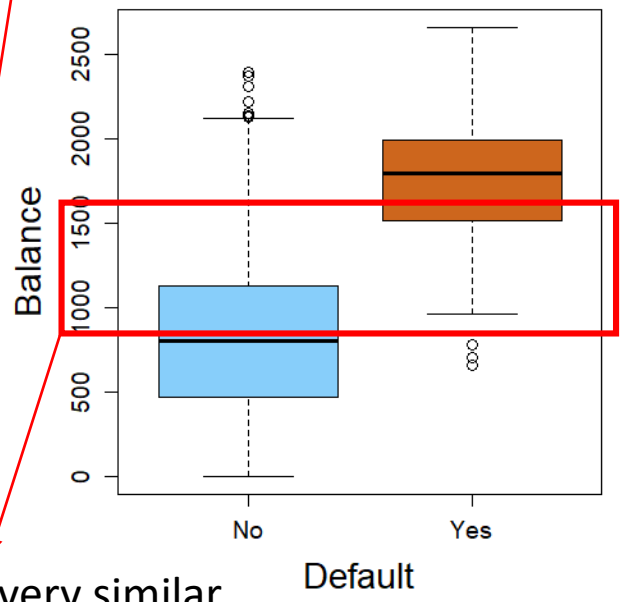












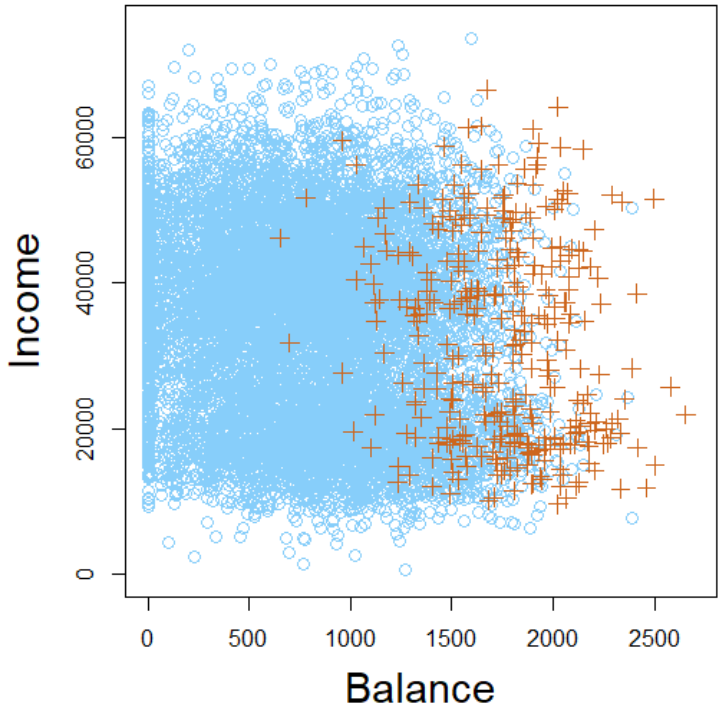
% of defaulted indicates probability

Relevant areas look very similar
→ box plots cannot indicate difference

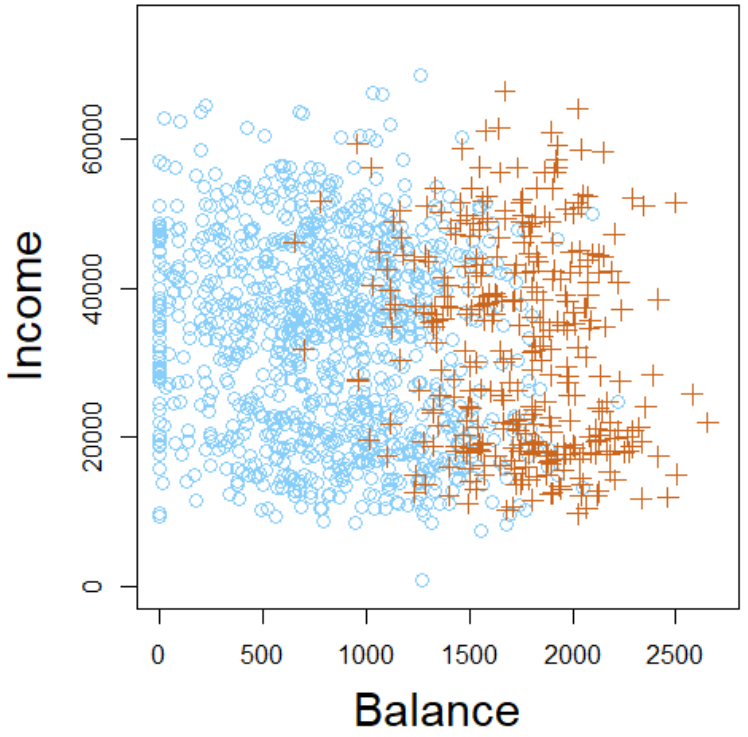
Quiz – www.menti.com, code 60 81 37

If you were the author of ISL, which plot would you use in the book?

A: the figure with all data shown



B: the figure showing fewer non-defaulted points



Logistic regression (cont.)

Plots and ethical aspects

Multivariate vs single variable case

Confounding

Recall: estimating probability

- The logistic regression equation estimates the log-odds by plugging in the coefficient estimates:

$$\log\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

- The estimated probability can be computed from this equation:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_p X_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\cdots+\hat{\beta}_p X_p}}$$

Student as single predictor

- Student is a categorical variable indicating student status

```
Call:
glm(formula = default ~ student, family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2970  -0.2970  -0.2434  -0.2434   2.6585

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

- R gives coefficient estimates → predicted default probability:

$$\hat{p}(X) = \frac{e^{-3.504+0.405\times\text{studentYes}}}{1+e^{-3.504+0.405\times\text{studentYes}}} = \begin{cases} \frac{e^{-3.099}}{1+e^{-3.099}} = 0.0431 = 4.31\% & \text{for students (studentYes=1)} \\ \frac{e^{-3.504}}{1+e^{-3.504}} = 0.0292 = 2.92\% & \text{for non-students (studentYes=0)} \end{cases}$$

Which variables to consider?

- Perform a variable selection procedure to get a good model

- Consider backward selection – start with model with all predictors:

```
Call:
glm(formula = default ~ student + income + balance, family = "binomial",
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
income       3.033e-06  8.203e-06   0.370  0.71152
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Income is not significant
(in the presence of balance and student status,
income does not help to predict default);

Income has highest p-value among non-significant predictors (as it is the only non-significant one)
→ we drop Income, see next slide for resulting model

Model with student and balance

- The two-variable model after dropping income:

```
Call:
glm(formula = default ~ student + balance, family = "binomial"
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4578  -0.1422  -0.0559  -0.0203   3.7435

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- All predictors are highly significant → this is the final model; predicted default probability:

$$\hat{p}(X) = \frac{e^{-10.75-0.715\times\text{studentYes}+0.0057\times\text{balance}}}{1+e^{-10.75-0.715\times\text{studentYes}+0.0057\times\text{balance}}} = \begin{cases} \dfrac{e^{-11.465+0.0057\times\text{balance}}}{1+e^{-11.465+0.0057\times\text{balance}}} & \text{for students (studentYes=1)} \\ \dfrac{e^{-10.75+0.0057\times\text{balance}}}{1+e^{-10.75+0.0057\times\text{balance}}} & \text{for non-students (studentYes=0)} \end{cases}$$

Logistic regression (cont.)

Plots and ethical aspects

Multivariate vs single variable case

Confounding

Coefficient of student

Outputs for student only & student+balance models:

```
Call:
glm(formula = default ~ student, family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2970  -0.2970  -0.2434  -0.2434   2.6585

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

```
Call:
glm(formula = default ~ student + balance, family = "binomial"
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4578  -0.1422  -0.0559  -0.0203   3.7435

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

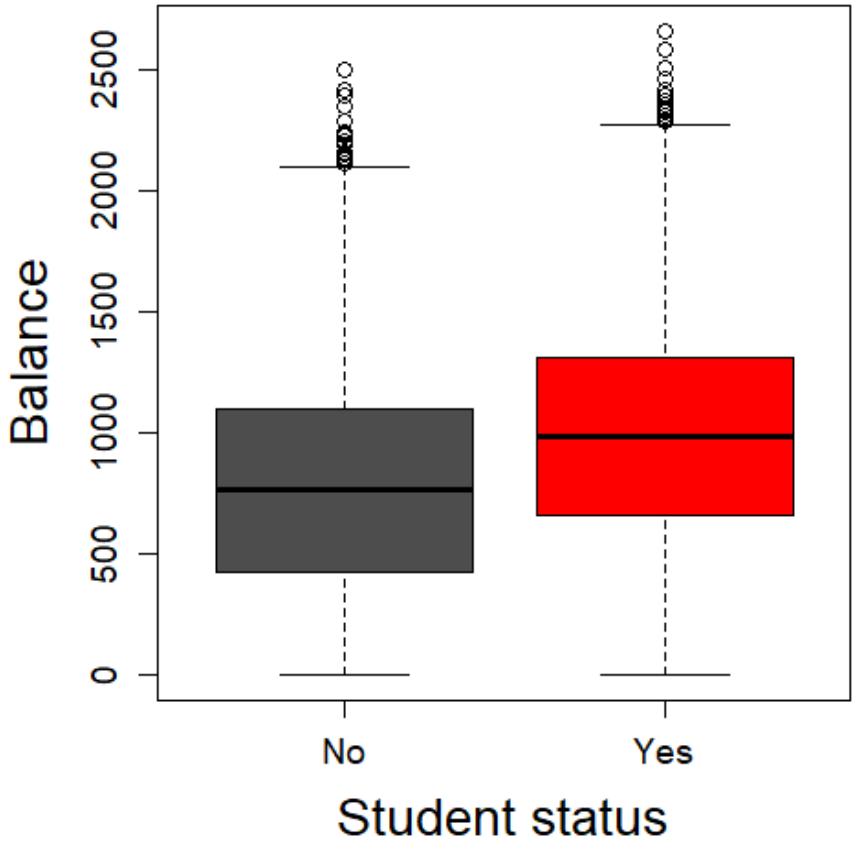
> 0 in single variable model,
< 0 in multivariate model

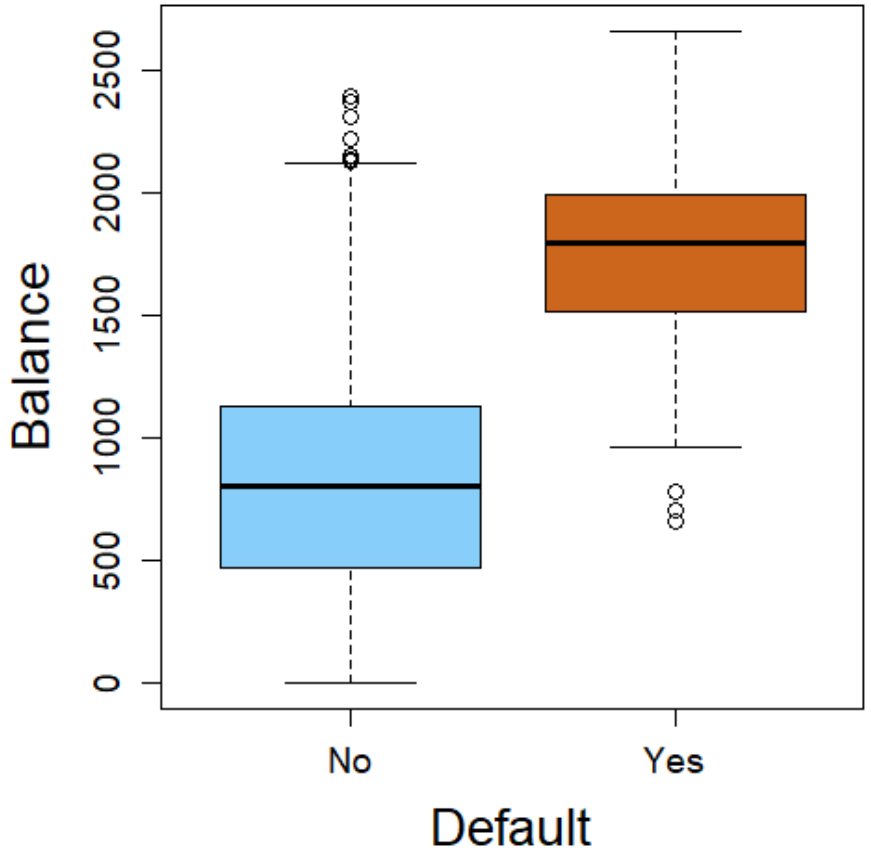
Are students more likely or less likely to default??

- Without further information → students have higher probability
- Among people with given balance → students have lower probability

How is this possible?

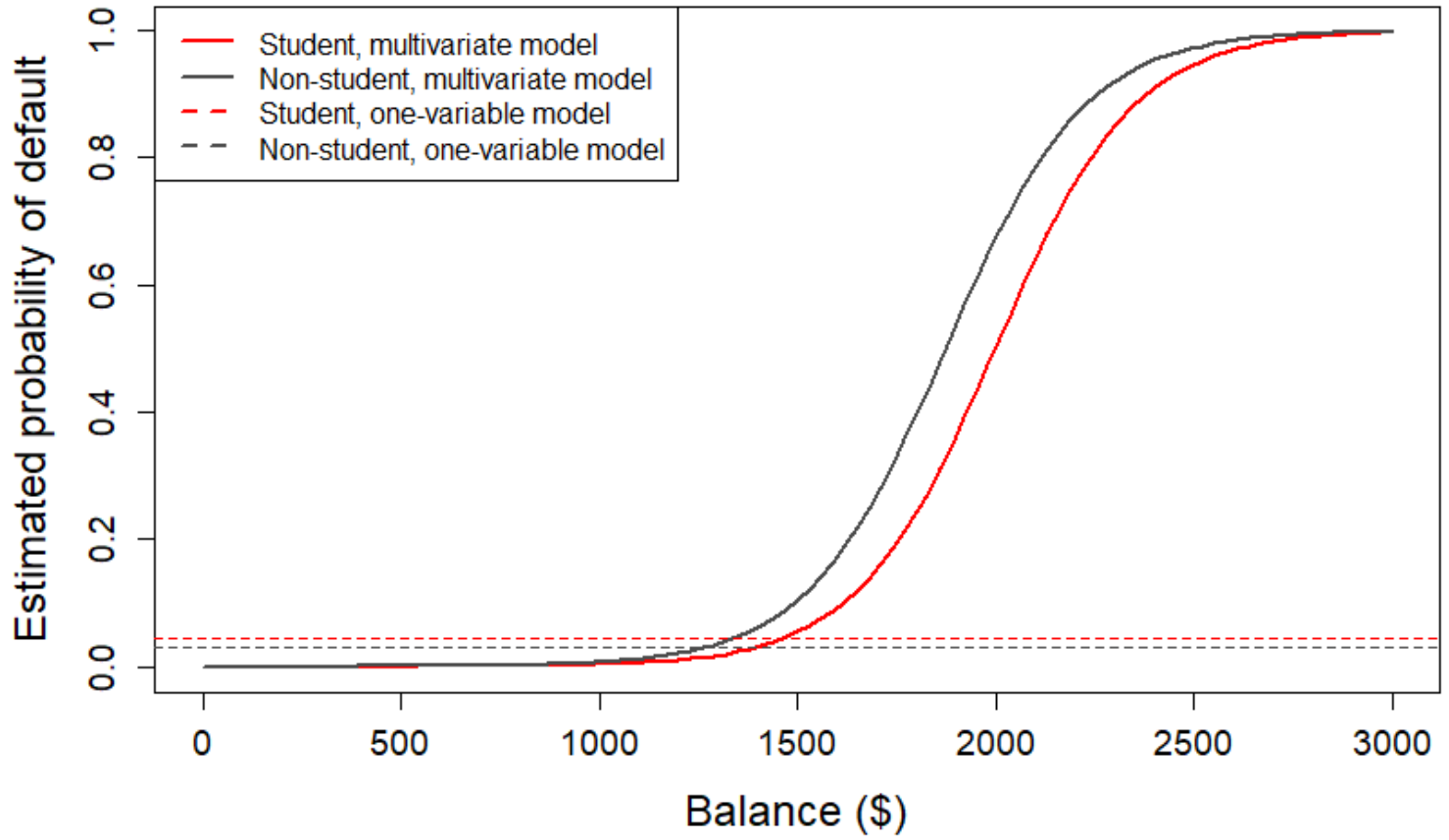
Being a student is associated with higher credit card balance which is associated with higher probability of default





Definition of confounding

Confounding is a phenomenon when the effect of a predictor on the response is turned around by another variable (→ results in multivariate model are very different from one-variable model)



Multinomial logistic regression

More than 2 response classes

We may want to classify a response variable with >2 classes:

- Select most likely diagnosis based on symptoms (epileptic seizure / drug overdose / stroke)
- Predict most likely outcome of car crash at given speed (fatality / serious injury / slight injury / no injury)
- Select preferable transport mode for goods (air transport / road transport / rail transport / water transport)

Multinormal logistic regression equation

- There are different forms of this model

- In the equation below (also used in the R package glmnet), each class gets its own linear model:

$$\Pr(Y = k|X) = \frac{e^{\beta_{0,k}+\beta_{1,k}X_1+\beta_{2,k}X_2+\cdots+\beta_{p,k}X_p}}{\sum_{m=1}^{K} e^{\beta_{0,m}+\beta_{1,m}X_1+\beta_{2,m}X_2+\cdots+\beta_{p,m}X_p}}$$

- We will see an example in the computer lab

Feed-forward

Feed-forward quiz

We will review the course material in w7. Which parts should we emphasize more?

1. Go to www.menti.com
2. Enter the code 38 50 70
3. Answer the questions or enter other comments related to the course or today's lecture