Statistical modeling in logistics

MMS075

Lecture 5b – 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

- Preparation for logistic regression
  - The classification problem
  - Reviewing the exponential function and logarithm
- Logistic regression
  - Motivation of the equation
  - Interpretation of the coefficients
  - Making predictions
- Feedback

Recommended resources

Reading in ISL: Ch 4 introduction and sections 4.1-4.3 for theory, 4.6.1, 4.6.2 for R codes

Online resources for exponential function and logarithm (examples):

Nykamp DQ, "Basic idea and rules for logarithms." From Math Insight. https://mathinsight.org/logarithm_basics

https://www.mathsisfun.com/algebra/exponents-logarithms.html

http://tutorial.math.lamar.edu/Classes/CalcI/ExpLogEqns.aspx

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

- **Introduction to Classification** (10:25)
- **Logistic Regression and Maximum Likelihood** (9:07)
- **Lab: Logistic Regression** (10:14)

Preparation for logistic regression

The classification problem

Reviewing the exponential function and logarithm

Qualitative response

So far we have considered:

- Quantitative response & quantitative predictors
- Quantitative response & categorical predictors (→ defined dummy variables to deal with those)

What to do if the **response is categorical**? Examples:

- Select most likely diagnosis based on symptoms
- Classify email as regular email or spam
- Predict most likely outcome of car crash at given speed

Logistic regression equation

- Logistic regression is very commonly used to address classification problems

- The model equation for binary response is provided below, motivation & properties will be discussed in later slides

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

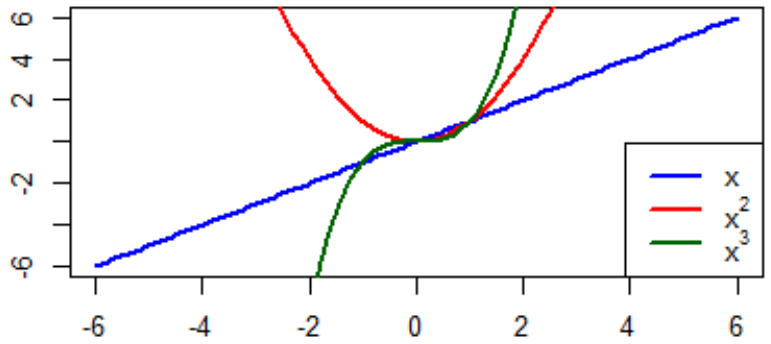
- Note: in line with the ISL book, we use ”log()” and not ”ln()” to denote natural logarithm (i.e. with base $e = 2.718...$)

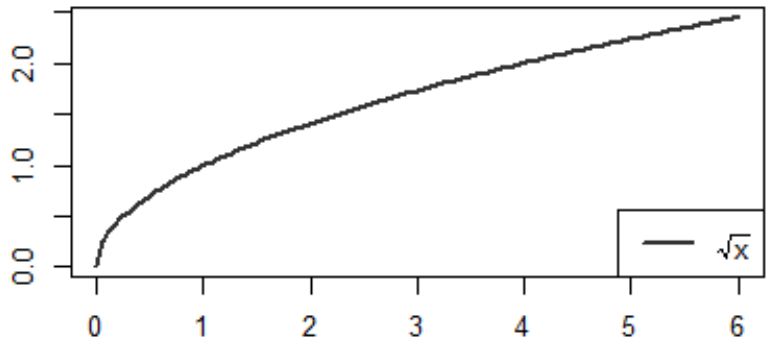
Preparation for logistic regression

The classification problem

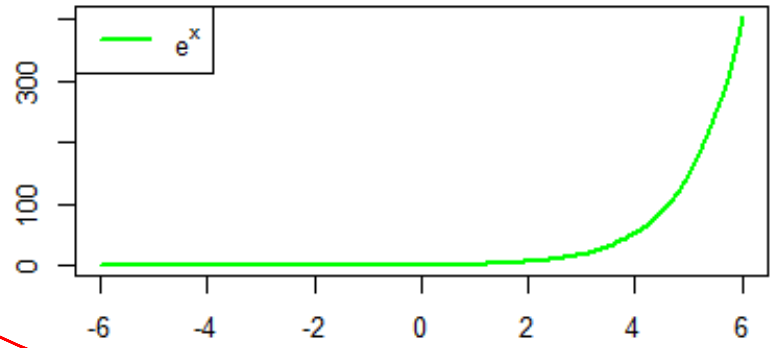
Reviewing the exponential function and logarithm

Recall: graphs of some important functions

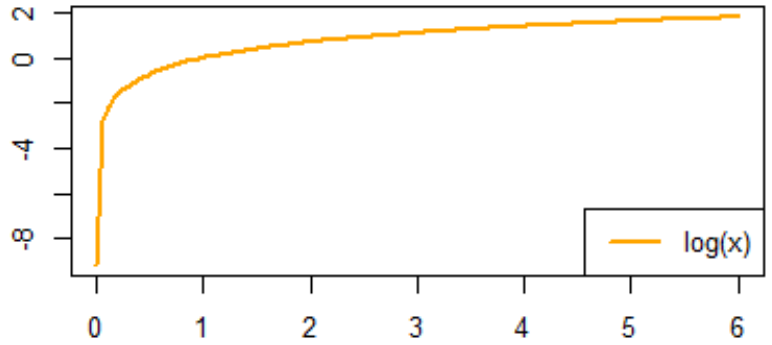




Exponential function



Logarithmic function



Basic properties of exp() and log()

- Some important properties are described below; detailed descriptions and exercises are available online (e.g. at the sites listed among recommended resources)

- $\exp(x) = e^x$ is defined for all values of x and is always positive

- $\log(x)$ is defined for positive values of x as the inverse of e^x; that is, $\log(x)$ is the number such that $e^{\log(x)} = x$

- Both functions are strictly increasing in x (i.e. take larger values when increasing x)

Basic properties of exp() and log() cont.

- $e^0 = 1, \log(1) = 0$

- exp(x) > 1 for x > 0 and exp(x) < 1 for x < 0 → log(x) > 0 for x > 1 and log(x) < 0 for 0 < x < 1

- Basic operations with the exponential function and logarithm:

$$e^{a+b} = e^a \cdot e^b \qquad \log(a \cdot b) = \log a + \log b$$

$$e^{a-b} = \frac{e^a}{e^b} \qquad \log(\frac{a}{b}) = \log a - \log b$$

$$e^{a \cdot b} = (e^a)^b \qquad \log a^b = b \cdot \log a$$

This equation was not included in Lecture 5a, and also the second bullet has been extended and rewritten

Logistic regression

Motivation of the equation

Interpretation of the coefficients

Making predictions

How to address categorical response?

- Why can't we give the responses numerical values and use linear regression to predict those?

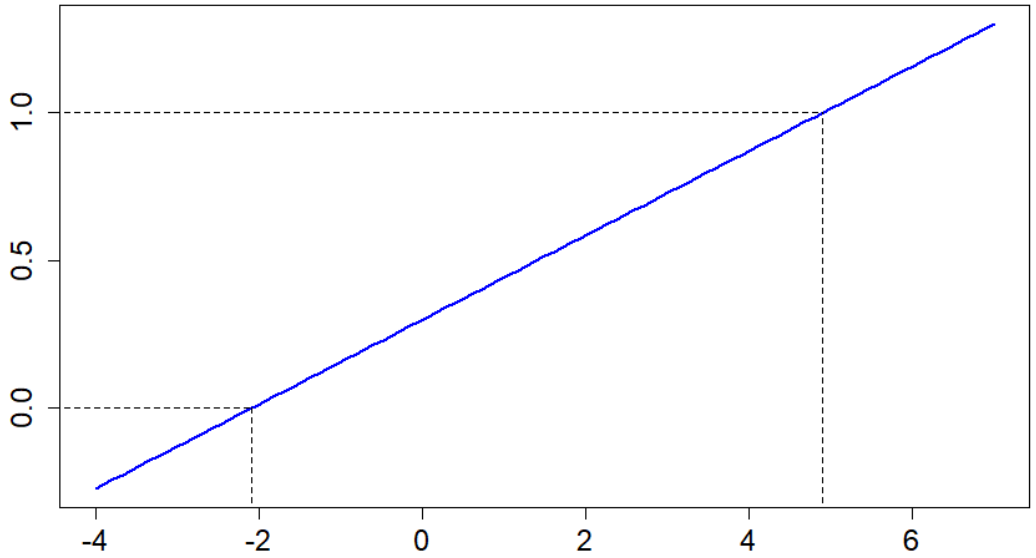
- The response categories are not ordered → it would be WRONG to randomly assign values 0, 1, 2, ..., because you would then create relationships between categories that are not real and don't make sense (e.g. epileptic seizure = 2x drug overdose)

$$Y = \begin{cases} 2 \text{ if the diagnosis is epileptic seizure} \\ 1 \text{ if the diagnosis is drug overdose} \\ 0 \text{ if the diagnosis is stroke} \end{cases}$$

But we can only predict numbers...

- How to relate an outcome to a number that we can then predict? Idea: look at the probability of the outcome!

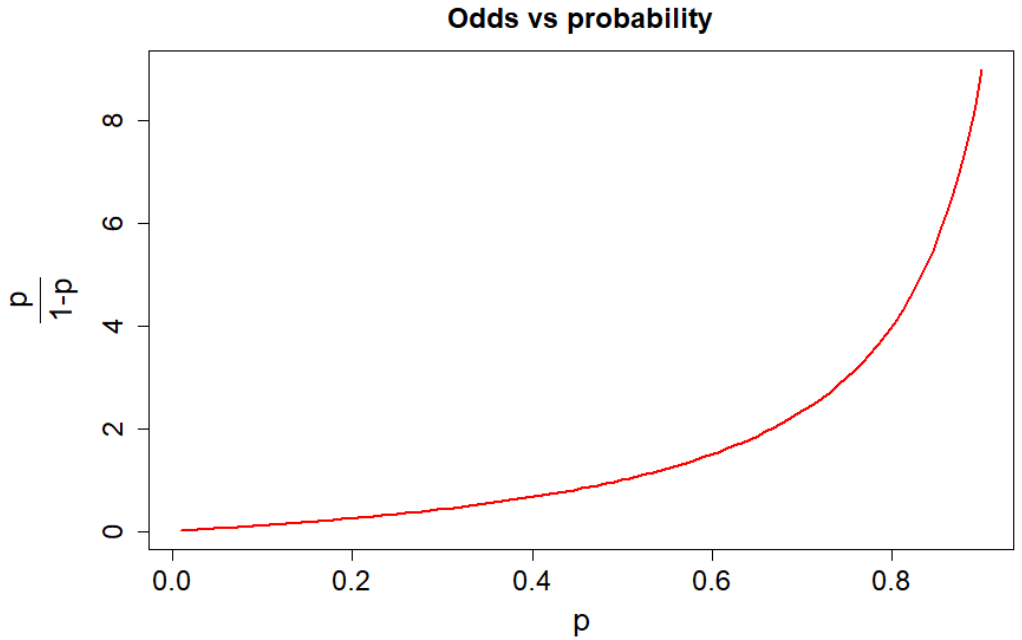
- Problem: linear regression gives predictions from -∞ to ∞, i.e. numbers of all sizes while probabilities are between 0 and 1



Transform probability to an infinite range

- Transformation 1: instead of the probability p of the outcome, model the **odds**, i.e. p/(1-p)

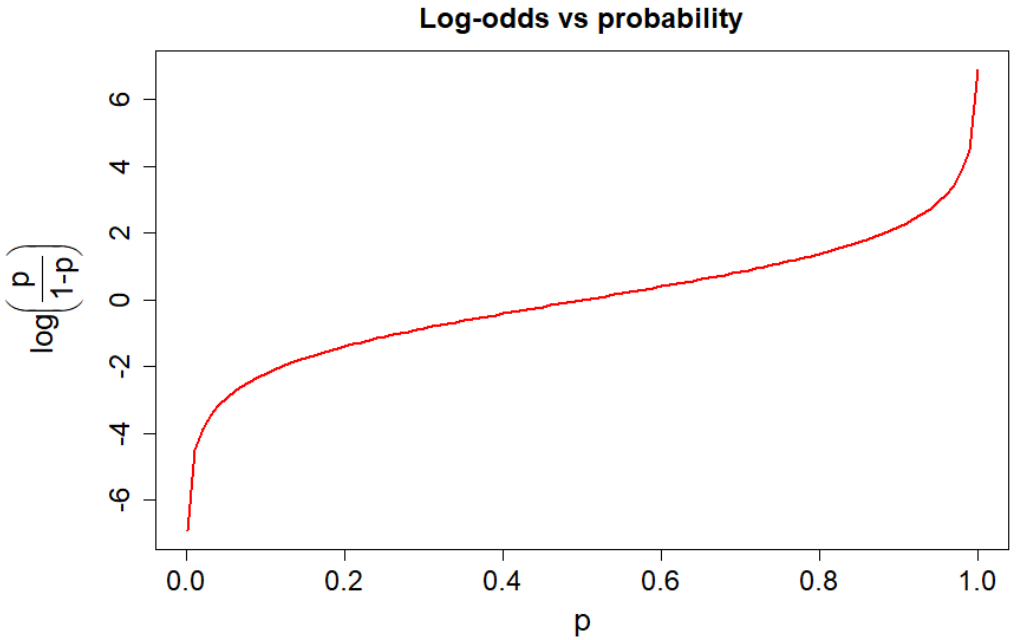
- Odds has a range from 0 to ∞ → this is not enough



Take logarithm to get <0 values as well

- Transformation 2: Consider the **logarithm of odds** (also called **log-odds** or **logit**) of the outcome

- The log-odds log(p/(1-p)) takes all values from -∞ to ∞ → we can use linear regression to predict it



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:

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

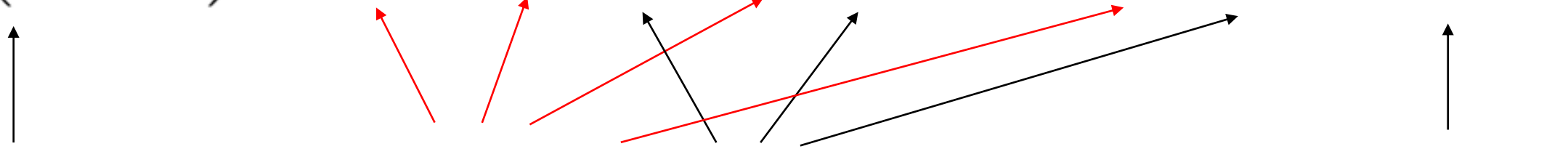
For given values of all predictors: $X = (X_1, X_2, \ldots, X_p)$, we denote **the probability of a case** by $p(X)$, i.e. we define the following conditional probability:

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

Logistic regression model for binary response

- Log-odds of ”success” 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 (with L>2 levels), this needs to be modified

Logistic regression

Motivation of the equation

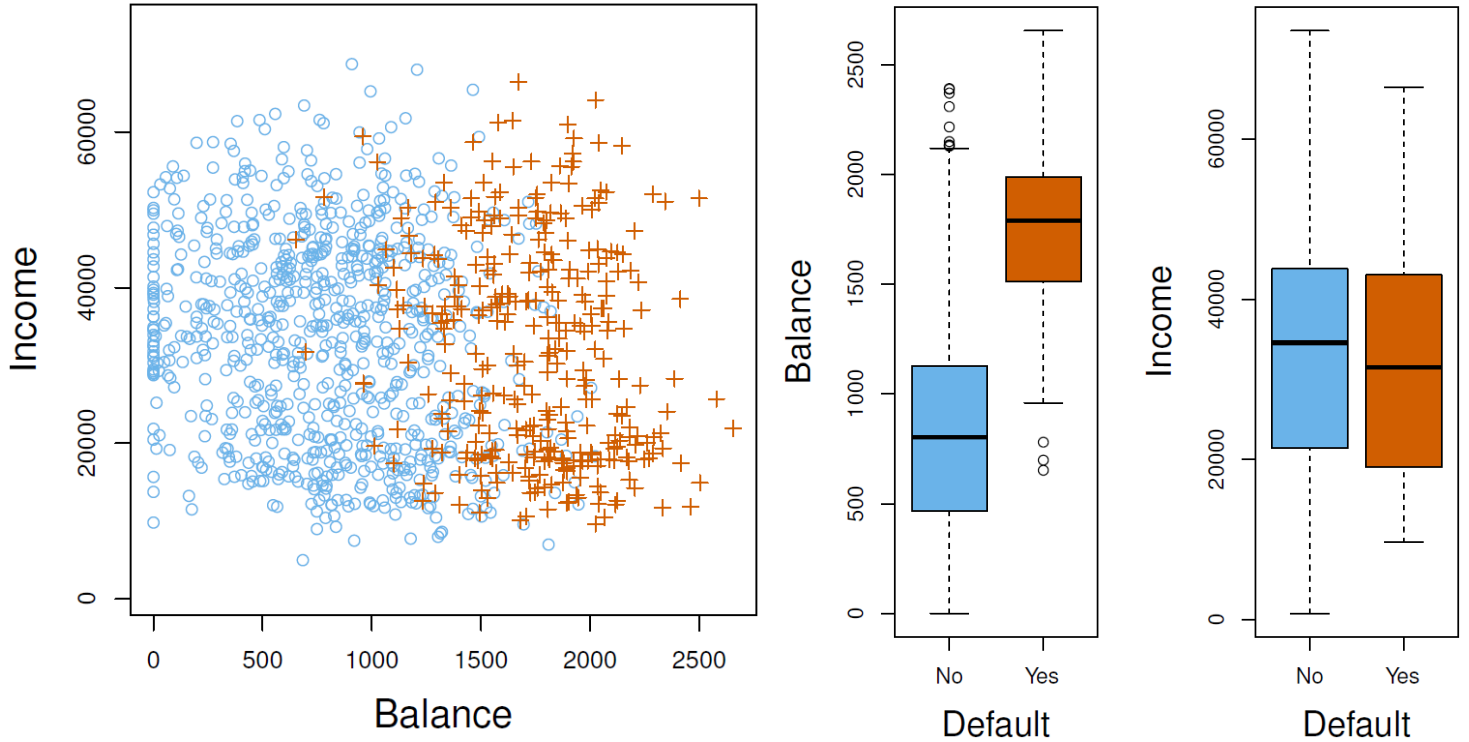
Interpretation of the coefficients

Making predictions

Recall credit card default example in ISL

- How can a bank predict if an individual will default on their credit card payment (i.e. fail to pay the credit card debt by the due date)?

- Predictors considered: annual income & monthly credit card balance



Source:
Figure 4.1 in ISL

Using only 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
```

Interpretation is in terms of (log-)odds

- The coefficients give how to estimate the log-odds of default based on balance:

$$\log\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right) = \hat{\beta}_0 + \hat{\beta}_1 \times \text{balance}$$
$$= -10.65 + 0.0055 \times \text{balance}$$

- As usual: $\hat{\beta}_1 = 0.0055$ → **log-odds of default increases by 0.0055** from a one-unit increase in balance

- Remembering that $e^{a+b} = e^a \cdot e^b$, this means that the **odds of default is multiplied by $e^{0.0055} = 1.005515$** from a one-unit increase in balance

- The relationship with the probability is non-linear → we can only say that the **probability of default increases** when balance increases

Logistic regression

Motivation of the equation
Interpretation of the coefficients
Making predictions

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:

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

Credit card default example

- We know from software output that:

$$\log\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right) = -10.65+0.0055\times\text{balance}$$

- The estimated probability of default for given values of balance is as specified on previous slide:

$$\hat{\Pr}(\text{default}=\text{Yes}|\text{balance}) = \hat{p}(X) = \frac{e^{-10.65+0.0055\times\text{balance}}}{1+e^{-10.65+0.0055\times\text{balance}}}$$

- For example, the estimated probability of default for a person with balance = $1000 is:

$$\hat{p}(X) = \frac{e^{-10.65+0.0055\times 1000}}{1+e^{-10.65+0.0055\times 1000}} = \frac{e^{-5.15}}{1+e^{-5.15}} = 0.0058 = 0.58\%$$

Feedback

Student representatives

- The names of student representatives for this course & contact information are published on the course homepage
- Please inform the student representatives about your impression of the course so far, e.g.:
  - Overall opinion
  - Potential issues
  - Improvement suggestions
- This is important even if you give feedback via www.menti.com

Feedback quiz - optional

Feedback is essential to me so that I can improve the lectures during the course. All comments about today's class or the course in general are welcome!

If you are willing to give feedback, please follow these steps:

1. Go to www.menti.com
2. Enter the code 32 21 69
3. Answer the questions or enter other comments related to the course