Analysing Social Science Data Using R

Class 11

Generalized Linear Models: Logistic Regression

Michał Bojanowski[1] mbojan@icm.edu.pl
Zbigniew Karpiński[2] zkarpinski@ifispan.waw.pl

[1]ICM University of Warsaw

[2]IFiS PAN

April 23, 2012
SNS, Warsaw

Outline

1 Linear Regression

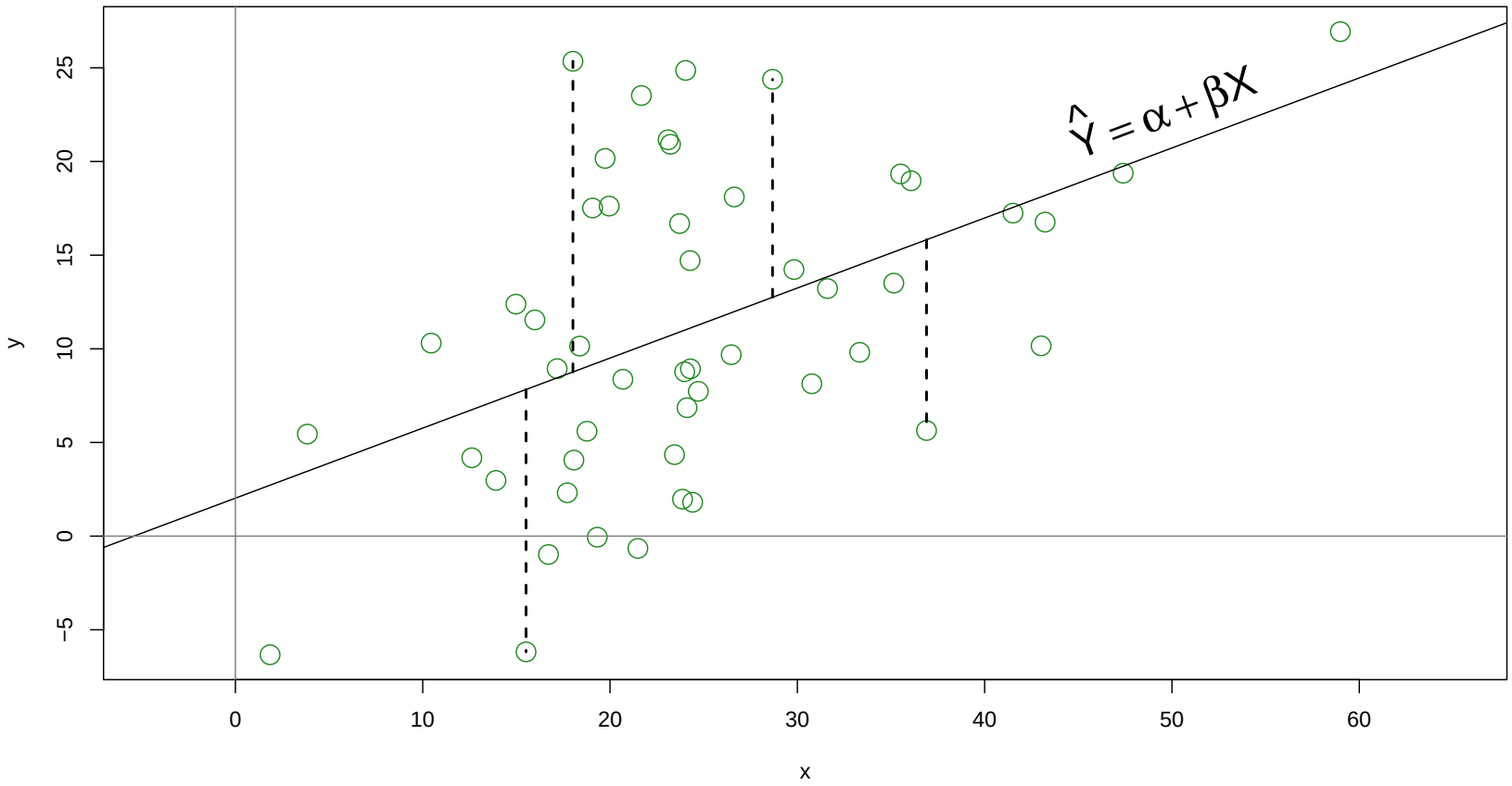
2 Generalized Linear Model

3 Logistic regression

Linear regression

- Requires continuous (interval or ratio scale) dependent variable.
- Requires continuous or dummy independent variables.
- Predicting values of dependent variable using a linear function of independent variables.
- Estimated parameters:
  - Intercept (α): What's the value of Y if all the Xs are equal to 0.
  - Slope (β): how much Y increases if X increases by 1 and all other variables are held constant.

Linear regression graphically



Generalized Linear Model (GLM)

- GLMs extend Linear Regression to settings where the dependent variable can be non-continuous. Examples:

Counts	Poisson regression, Negative binomial, Log-linear models.
Binary	Logistic regression, Probit regression.
Categorical	Multinomial logit.
Ordinal	Ordered logit, ordered probit
...	

GLM defined

$$g(Y) = \alpha + \beta X$$

Y Dependent variable

X Independent variable

$g(\cdot)$ Link function.

α, β Regression coefficients.

- Right-hand side is identical to linear regression
- Choice of $g(\cdot)$ and probability distribution for Y determines the type of model.

Logistic regression

Logistic regression is a GLM, in which:

1. Y **is a binary variable** (e.g. "success" or "failure") occurring with an unknown probability p
2. Unknown probability p is a function of independent variables via **logit link**:

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

Logit and odds

If p is a **probability** of a particular event, then **odds** are $\frac{p}{1-p}$.
Odds are an alternative language to talk about chances of events:
Examples:

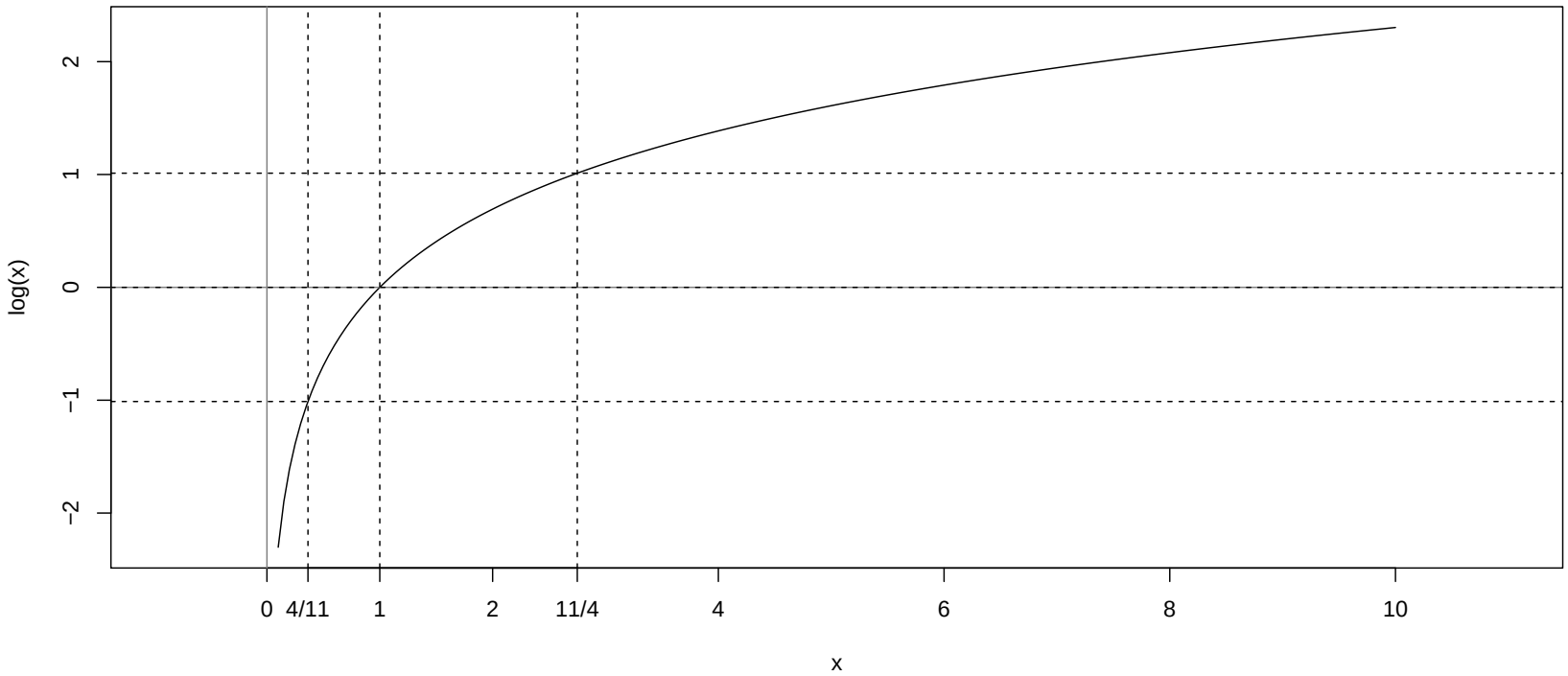
- Tossing a coin will give head or tail with prob. 0.5 each. Odds are $\frac{0.5}{0.5} = \frac{1}{1}$ or simply 1.
- According to bookmakers, Spain is likely to win Euro 2012 with odds $\frac{11}{4}$.

$$\frac{p}{1-p} = d \qquad \leftrightarrow \qquad p = \frac{d}{1+d}$$

So the probability for Spain to win is $\frac{11}{15}$.

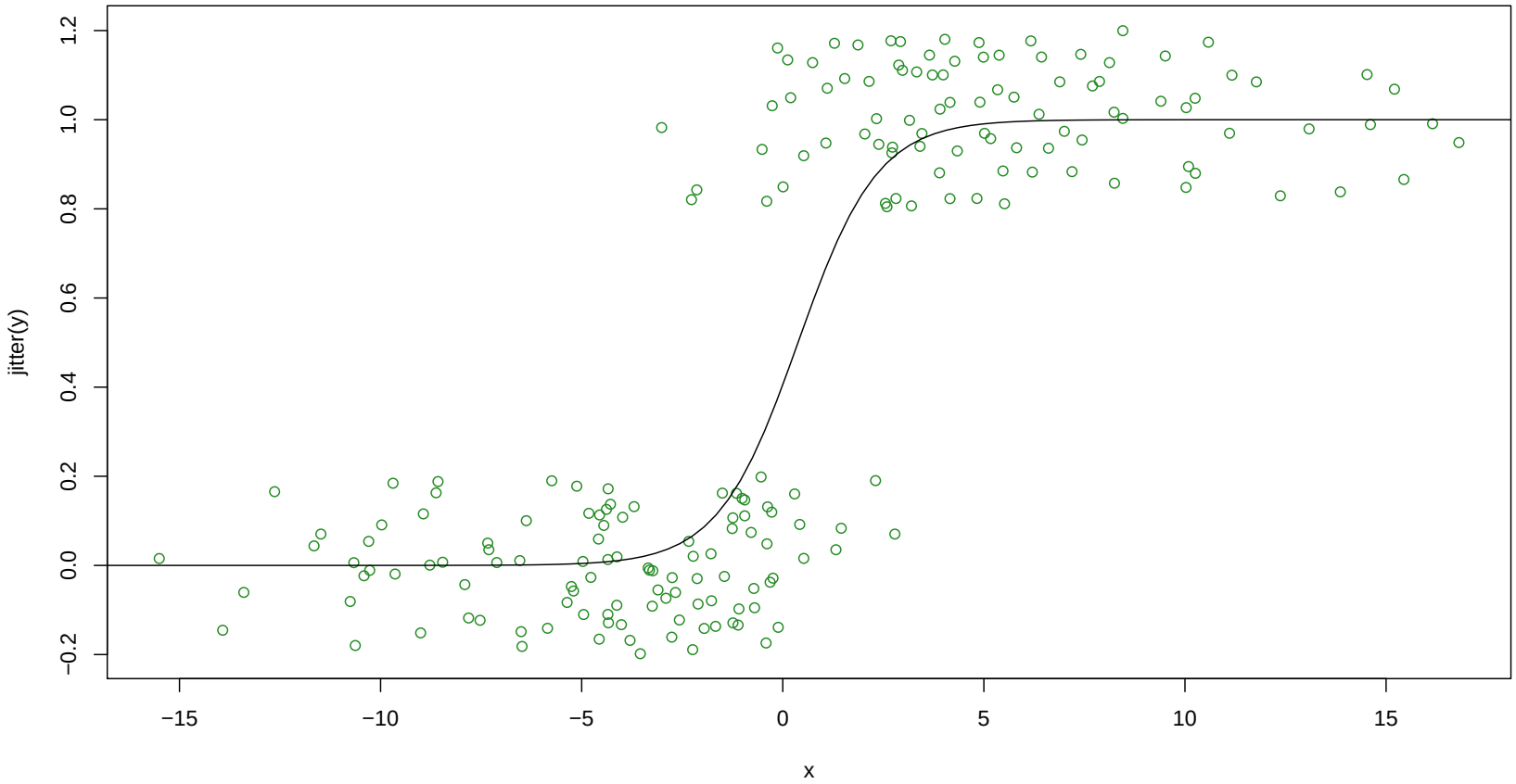
Logit and odds: why log?

To make it symmetric around 0



$$\log\left(\frac{p}{1-p}\right) = -\log\left(\frac{1-p}{p}\right)$$

Logistic regression graphically



Interpreting coefficients

Exponentiating to get rid of the log:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X \qquad \leftrightarrow \qquad \frac{p}{1-p} = e^{\alpha} e^{\beta X}$$

In particular:
If $X = 0$:

$$\frac{p}{1-p} = e^{\alpha}$$

If $X = 1$:

$$\frac{p}{1-p} = e^{\alpha} e^{\beta}$$

- e^{α} are the odds of $Y = 1$ if $X = 0$
- e^{β} is how much **times** the odds for $Y = 1$ will increase if X is increased by one.

Logistic regression in R

- Using function `glm`. Arguments identical to `lm`
- Additional argument required: `family=binomial("logit")`.

For example:

```
> mod <- glm( y ~ x1 + x2, data=df, family=binomial("logit"))
```

Use `exp` to transform β to e^{β}:

```
> 11/4
[1] 2.75
> log(11/4)
[1] 1.011601
> exp( log(11/4) )
[1] 2.75
```