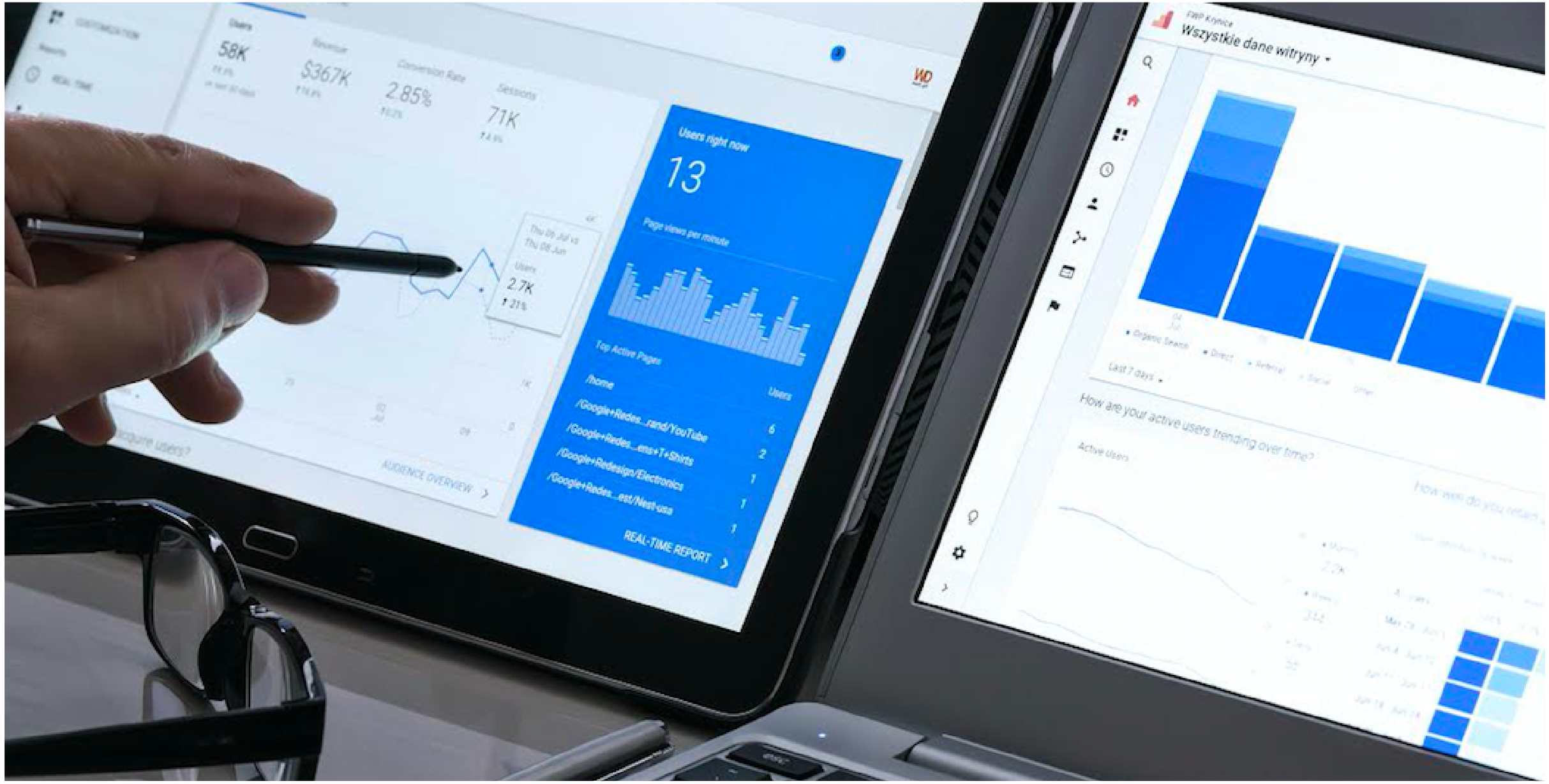
Overview of logistic regression

GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor

datacamp


58K
$367K
2.85%
71K
Users right now
13
Page views per minute
Top Active Pages
/home
Users
REAL-TIME REPORT
Wszystkie dane witryny
How are your active users trending over time?
Active Users

Chapter overview

- Overview of logistic regression
- Inputs for logistic regression in R
- Link functions

Why use logistic regression?

- Binary data: (0/1)
- Survival data: Alive/dead
- Choices or behavior: Yes/No, Coke/Pepsi, etc.
- Result: Pass/fail, Heads/tails, Win/lose etc.

What is logistic regression?

Default GLM for binomial family

Model of binary data

$$Y = \text{Binomial}(p)$$

Linked to linear equation

$$\text{logit}(p) = \beta_0 + \beta_1 x + \epsilon$$

Logit function

Logit defined as

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

Inverse logit defined as

$$\text{logit}^{-1}(x) = \frac{1}{1+\exp(-x)}$$

How to run logistic regression

Function:

```
glm(y ~ x, data = dat, family = 'binomial')
```

Inputs:

```
y = c(0, 1, 0, 0, 1...)
y = c("yes", "no"...)
y = c("win", "lose"...)
# Or any 2-level factor
```

Riding the bus?

- What makes people more likely to commute using a bus?
- Ride bus: **yes**, Not-ride bus **no**
- Do number of commuting days change the chance of riding the bus?
- 2015 commuter **data from Pittsburgh, PA, USA**

```
  CommuteDays Bus
1           5 Yes
2           2  No
```

Let's practice!

GENERALIZED LINEAR MODELS IN R

Bernoulli versus binomial distribution

GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor

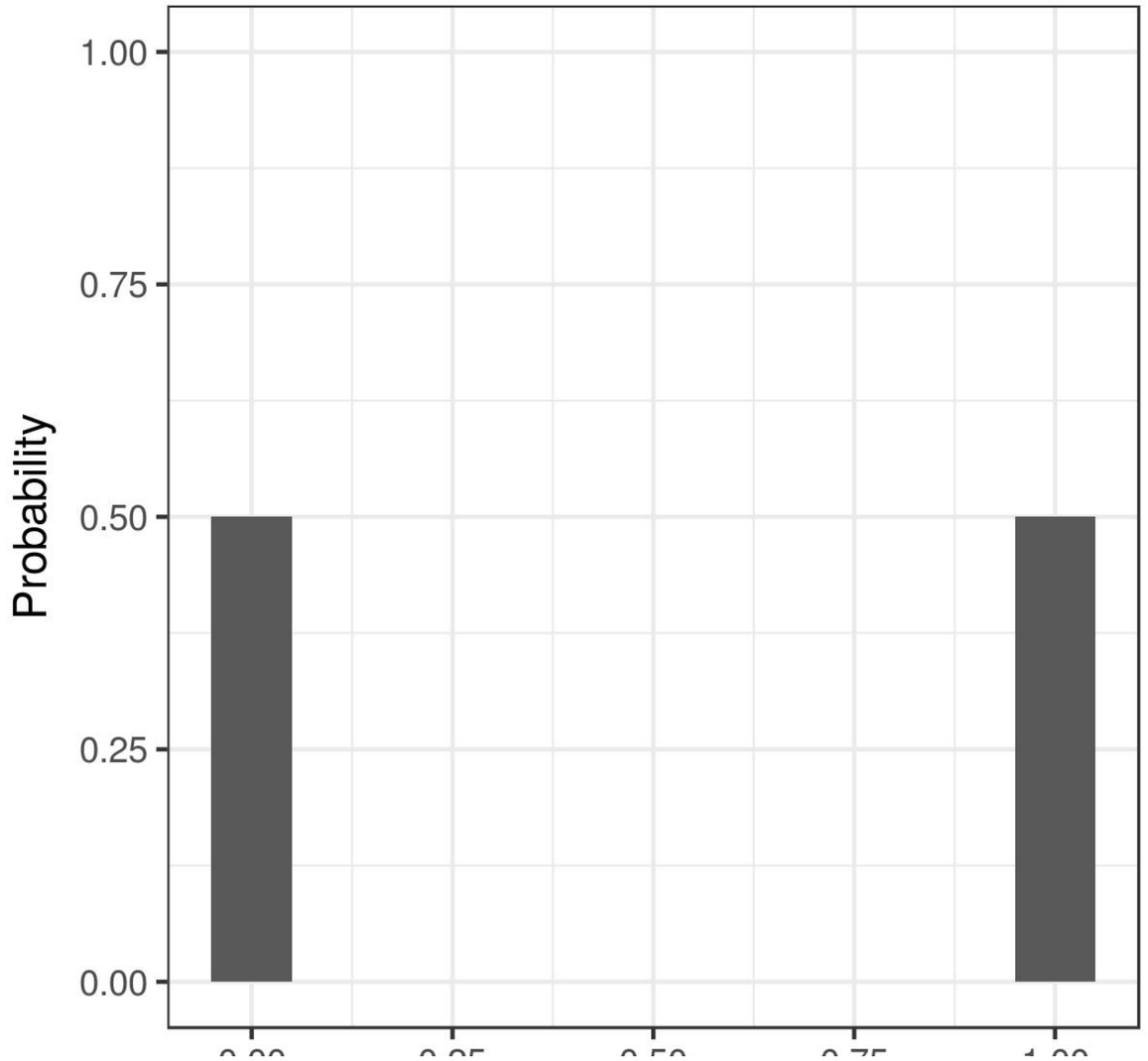
datacamp

Foundation of GLM

- Binomial and Bernoulli foundation of logistic regression
- Closely related to data input

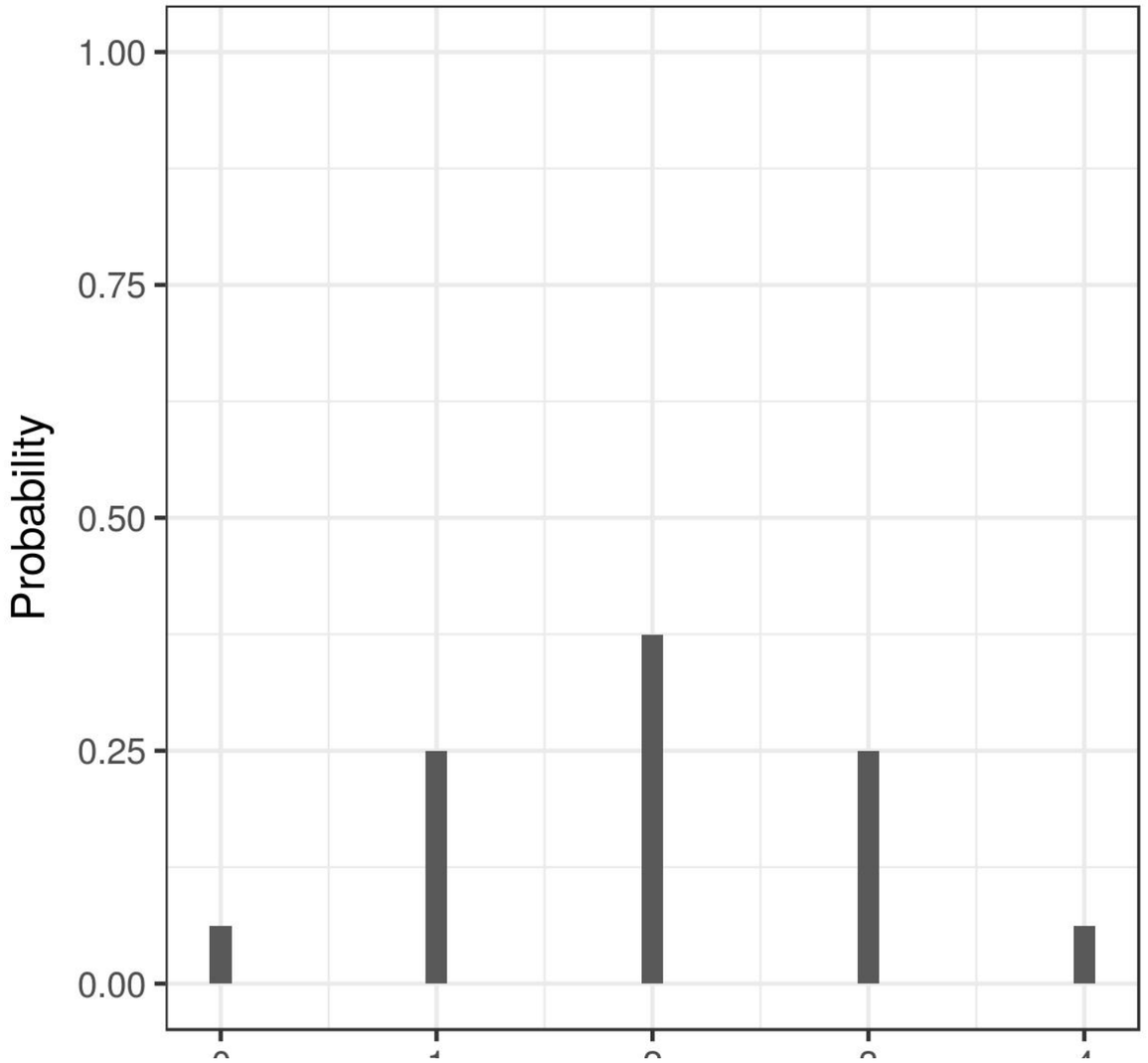
Bernoulli distribution

- Binary outcome: e.g., single coin flip
- Expected probability
  - k outcomes
  - with p probability
  - $f(k, p) = p^k(1-p)^{1-k}$
- Example of flipping 1 coin



Binomial distribution

- Discrete outcome: e.g., flipping multiple coins
- Expected probability
  - n trials
  - k outcomes
  - with p probability
  - $f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$
- Flipping 4 coins at once



Simulating in R

```
rbinom(n = , size = , p = )
```

- `n` : Number of random numbers to generate
- `size` : Number of trials
- `p` : Probability of "success"
- `size = 1` : Bernoulli

GLM inputs options

- Long format (Bernoulli format)
  - `y = c(0,1,...)`
  - Allows for variables for each observation
- Wide format (Binomial format)
  - Matrix: `cbind(success, failure)`
  - Proportion of success: `y = c(0.3, 0.1,...)` with `weights = c(1, 3, 2...)`
  - Looks at "groups" rather than individuals

Example

Long data:

- One entry per row
- Predictors for each response

```
response treatment  length
    dead         a 3.471006
    dead         a 3.704329
   alive         a 2.043244
   alive         b 1.667343
```

Wide data:

- One group per row
- Predictors for each group

```
group dead alive Total groupTemp
    a   12     2    14      high
    b    3    11    14       low
```

Which input method to use?

- What is your raw data structure? Long or wide?
- What variables do I have? Individual or group?
- Do want to make inferences about groups or individuals?

Let's practice!

GENERALIZED LINEAR MODELS IN R

Link functions- Probit compared to logit

GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor

datacamp

Why link functions?

- Understand and simulate GLMs
- Probit vs logit as example

Why probit?

- Demonstrate link function
- Used in some fields (e.g., toxicology)
- Preferred by some people

What is a probit?

- **Pro**bability un**it**
- Toxicology by Chester Bliss in 1934
- Computationally easier than logit
- Model know as probit analysis, probit regression, or probit model

Probit equation

Model of binary data

$$Y = \text{Binomial}(p)$$

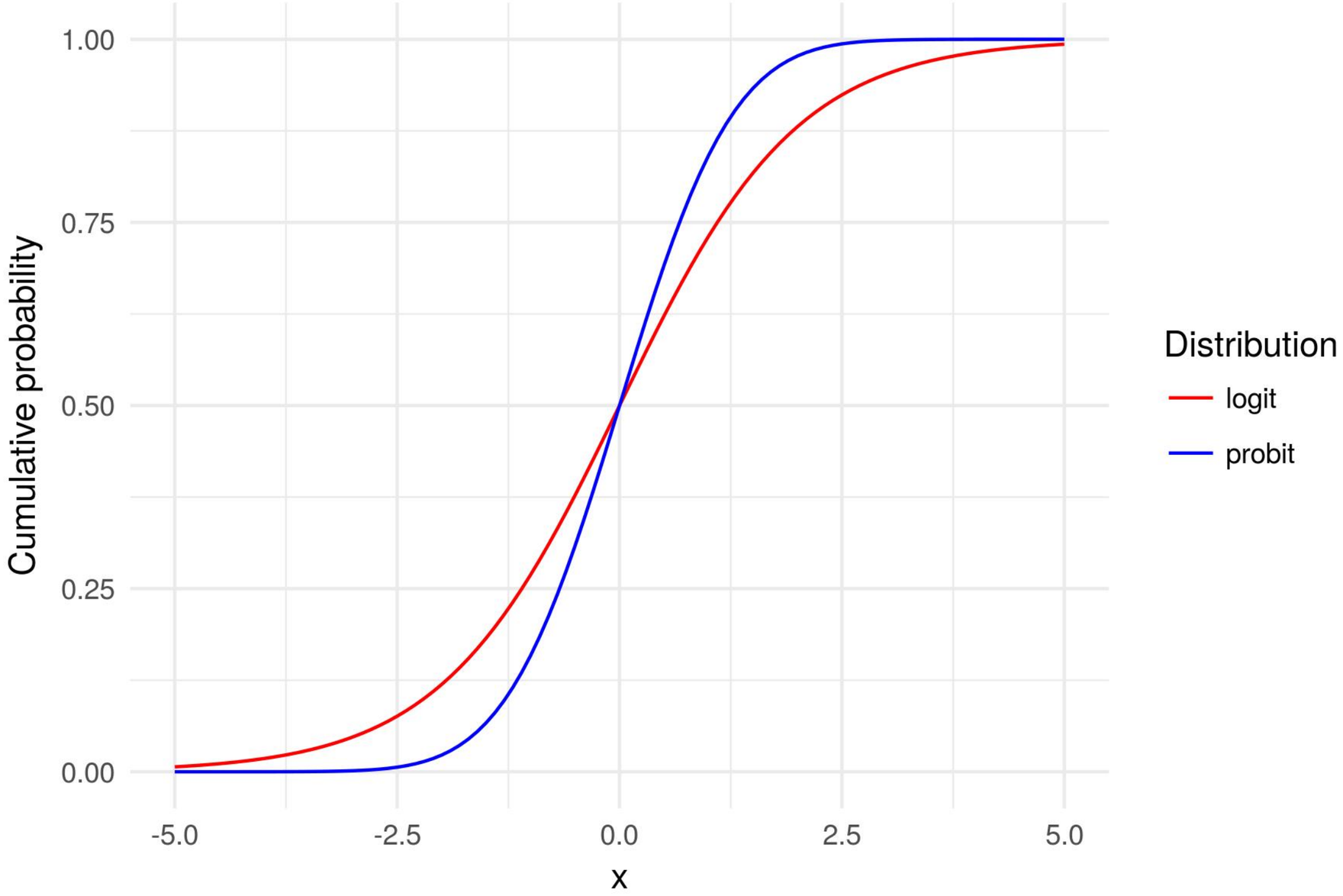
Linked to linear equation

$$\Phi^{-1}(p) = \beta_0 + \beta_1 x + \epsilon$$

Probit function

Based upon cumulative normal

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}z^2} \, dz$$


1.00
0.75
0.50
0.25
0.00
Cumulative probability
-5.0
-2.5
0.0
2.5
5.0
x
Distribution
logit
probit

Fitting a probit in R

- `family` option for `glm()`
  - Character: `glm(..., family = "binomial")`
  - Function: `glm(..., family = binomial()`)
- Default: `binomial(link = "logit")`
- Probit: `binomial(link = "probit")`
- Match instructions for DataCamp

Simulate with probit

Convert from probit scale to probability scale:

```
p = pnorm(-0.2)
```

Use probability with binomial distribution

```
rbinom(n = 10, size = 1, prob = p)
```

Simulate with logit

Convert from logit scale to probability scale:

```
p = plogis(-.2)
```

Use probability with a binomial distribution

```
rbinom(n = 10, size = 1, prob = p)
```

When to use probit vs logit?

- Largely domain specific
- Thicker tails of logit
- Either is tenable

Let's practice!

GENERALIZED LINEAR MODELS IN R