# Models of outcome and choice: The logit model

Silje Synnøve Lyder Hermansen

November 12, 2019

# Table of Contents

# Section 1

# A latent variable approach to GLMs

# Many outcomes are not continuous

**OLS assumes a continuous dependent variable. But many phenomena in the social sciences are not like that.**

- Vote choice, civil conflict onset, legislator performance, court rulings, time to compliance, etc.
- What phenomena are you interested in?

$\Rightarrow$ *OK. Let's strategize.*

# All regressions are linear(ized)

**The basic formulation in any regression describes a linear relationship between $x_i$ and $y_i$:**

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1}$$

- When $x_i$ increases with one unit, $y_i$ increases with $\beta$ units.
- If that relationship is not linear, we have to make it so:
  - by recoding the $x_i$
  - by recoding the $y_i \rightarrow$ we *linearize*.

# A latent variable

**A linear(ized) model requires a continuious dependent variable.**

- Imagine we are interested in unobservable variable, $z_i$, that describes our propensity towards something.
- Above a certain threshold ($\tau$) of $z_i$, observability kicks in and we can see $y_i$.
- The regression coefficients ($\beta$) in GLMs describe that relationship.

$\Rightarrow$ The latent variable approach is useful when interpreting the results.

# Example: The binomial model

**The logit model is a perfect example:**

$$y_i = \begin{cases} 1 & \Leftrightarrow \quad z_i > \tau \\ 0 & \Leftrightarrow \quad z_i \leqslant \tau \end{cases} \tag{2}$$

- The probability ($z_i$) of an outcome $y_i$ is continuous.
- Above a certain probability ($\tau$), we observe a positive outcome ($y_i = 1$).

$\Rightarrow$ *but how do we set the value of* $\tau$*?*

# From latent variable to descrete outcomes

**Statistical theory helps us describe how $z_i$ leads to $y_i$.**

- What kind of process generated our data? $\rightarrow$ data generating process (DGP)
- How can we best describe it? $\rightarrow$ choice of *probability distribution* (in GLM)

# The three components of GLMs

**When fitting the model, we need to make three choices:**

- A linear predictor: $\beta x_i$.
- A probability distribution: they're all in the exponential family
- A recoding strategy

# The three components of GLMs

**In R this translates to two additional arguments compared to your usual OLS.**

- A linear predictor: $\rightarrow$ (`y ~ x`).
- A probability distribution: $\rightarrow$ (`family =`)
- A recoding strategy $\rightarrow$ (`link = `).

```
glm(y ~ x, data = data, family = binomial(link = "logit"))
```

# Latent variable approach for interpretation

- The latent variable approach is useful when interpreting results.
- That's when we map *from* the latent variable *to* the observed outcome.

$\Rightarrow$ *When estimating the model, we have to go the other way 'round.*

# Table of Contents

# Section 2

# Recoding: How do we get from a binary to a continuous variable?

# Data structure

**We can only observe the outcome produced by the latent variable. There are two data structures for binary data:**

- classes of observations: e.g.: rats in a cage, coin tosses...
- case-based: e.g.: legislator votes, Brexit...

# Data structure

**We can only observe the outcome produced by the latent variable. There are two data structures for binary data:**

- classes of observations: e.g.: rats in a cage, coin tosses... $\rightarrow$ *the closest to the latent continuous variable.*
- case-based: e.g.: legislator votes, Brexit...

$\Rightarrow$ *we know the number of successes and trials in a cage/class/stratum. That's our starting point.*

# Let's start with the odds

**Despite binary outcomes, we want a continuous variable that is unbounded at both ends. We define a stratum and start comparing**:

- Odds: Compare number of successes with number of failures within a stratum$\rightarrow$ *continuous but highly skewed.*
- Logtransform the odds $\rightarrow$ *continuous and bell shaped.*

# Let's examplify with rats

**We kept a 1000 rats in a cage and a number of them died (failure) while others are still alive (success). How can we model this?**

# We calculate the odds

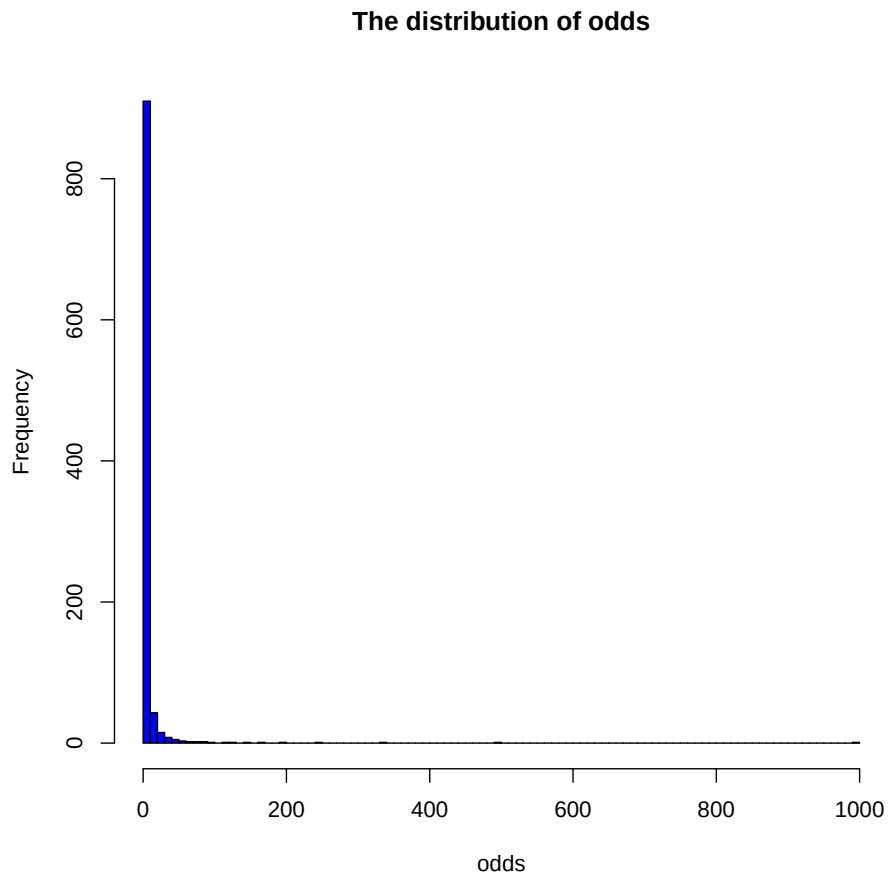
**We calculate the odds of surviving in a cage in a 1000 cages**

- Let's consider a series of 1000 trials where we let the successes go from complete failure (success = 0) to complete success (success = 1000)

```r
success = 0:1000
tries = 1000
#remember: failure = tries - success
odds <- success/(tries - success)

hist(odds, breaks = 100, col = "blue")

hist(log(odds), breaks = 101, col = "blue")

plot(log(odds), success, type = "l")
```

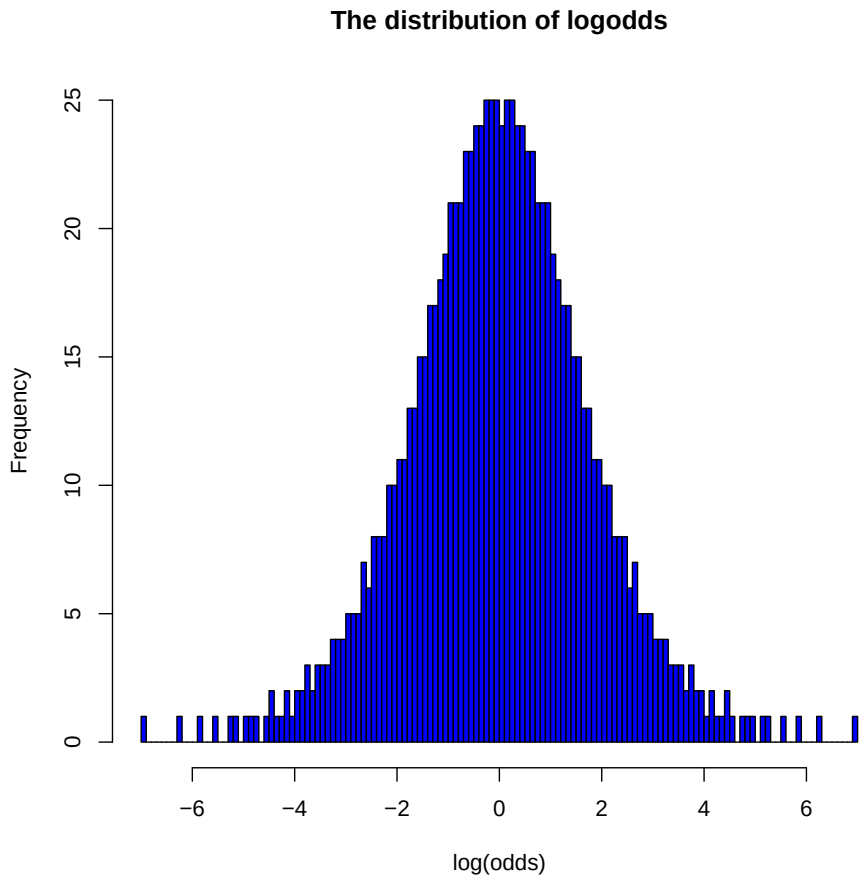
# Let's start with the odds

**We get a continuous but skewed variable.**



# Now, let's logtransform the odds

**We get a nice, bellshaped curve.**

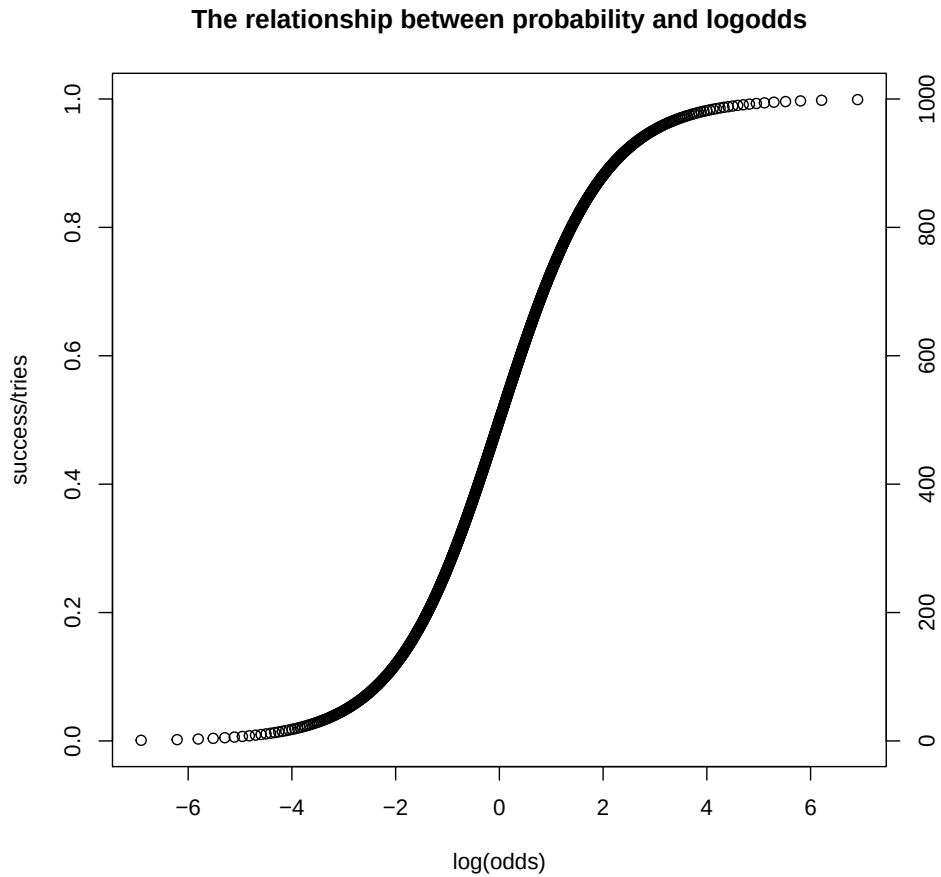


# Now, let's logtransform the odds

**This, we can run regressions on!**

# The famous S shape

**We can plot the logodds of success against the number of successes or their probability (it's the same).**



# Probability distributions for binary variables

**There are two, closely related probability distributions for binary outcomes**:

- The binomial distribution: $B(n, p)$
  - $p$ is the probability of success tells where on the x-axis (trials) the distribution is placed.
  - $n$ is the number of trials and defines the precision (width) of the distribution.
- The Bernoulli distribution: $Ber(p)$: when we only have only one trial.

# Individual-level outcomes

**In the social sciences, we usually look at case-level data**

- We rely on the odds (probability) of all cases with identical $x$-values to calculate $y$

# Recoding in 3 steps

**We can do the same using probabilities**:

- Calculate the proportion of successes: probability
- Calculate the odds of success from probability
- Logtransform the odds

# Recoding in 3 steps

**I'll say that again**:

proportion/probability $\rightarrow$ odds $\rightarrow$ logodds

# Table of Contents

# Section 3

## Why all the fuzz? Why not OLS?

# Distributions in OLS and maximum likelihood

- In OLS: The residuals must be normally distributed, but not the $y$
- In ML: The $z$ must follow a probability distribution, but residuals are fixed.

$\Rightarrow$ *This is due to how they are estimated*

# We want reliable standard errors

- In OLS: The residuals must be normally distributed, but not the $y$
- In ML: The $z$ must follow a probability distribution, but residuals are fixed.

$\Rightarrow$ *This is due to how they are estimated*

## Subsection 1

# Back and forth: Logistic and logit transformation

# The logit transformation

**When we go from outcomes to latent variable we use the logic transformation**.

$$logit(p) = log(\frac{p}{1-p}) \tag{3}$$

$\Rightarrow$ *This what R does when estimating our model*

# The logistic transformation

**When we go from the latent variable to outcomes we use the logistic transformation**.

$$logit^{-1}(logodds) = \frac{exp(logodds)}{1 + exp(logodds)} = \frac{1}{1 + exp(-logodds)} \quad (4)$$

$\Rightarrow$ *This what we do when interpreting our model*

# Table of Contents

# Section 4

## Interpretation: So... what did I find?