Models of outcome and choice: The logit model

Silje Synnøve Lyder Hermansen

```
##
## Vedhæfter pakke: 'dplyr'

## De følgende objekter er maskerede fra 'package:stats':
##
##     filter, lag

## De følgende objekter er maskerede fra 'package:base':
##
##     intersect, setdiff, setequal, union
```

Before we start

Before we start

Where are we?

Assumptions of the linear model

Linear models (OLS) rely on two assumptions that are often violated

- observations are independent and identically distributed (iid)
- **outcomes are continuous and unbounded** (next 7 weeks)

$\Rightarrow$ *this class: alternative models when these are not satisfied.*

Take 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.

$\Rightarrow$ *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:**

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- When x_i increases with one unit, y_i increases with β 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 continuous dependent variable.**
  - Imagine we are interested in an unobservable variable, z_i, that describes our propensity towards something.
  - Above a certain threshold (τ) of z_i, observability kicks in and we can see y_i.
  - The regression coefficients (β) in GLMs describe the $z \sim x$ relationship.

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

Example: The binomial model

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

$$y_i = \begin{cases} 1 & \text{if } z_i > \tau \\ 0 & \text{if } z_i \leq \tau \end{cases}$$

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

$\Rightarrow$ *But how do we set the value of τ?*

From latent variable to discrete 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: βx_i.
  - A probability distribution: they're all in the exponential family.
  - A recoding strategy.

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 =`).

The three components of GLMs

- **In R, this translates to two additional arguments compared to your usual OLS:**
  - A linear predictor: → (`y \sim x`).
  - A probability distribution: → (`family =`)
  - A recoding strategy → (`link =`).

```
# Example R code for a GLM model
mod <- 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.*

Take 2: Recoding from binary to continuous

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.*

The binomial distribution: successes and failures

The binomial distribution: successes and failures

How does the binomial distribution map descrete outcomes (0 or 1) to something continuous?

- let's start with the intercept-only model (no predictors, just a base-line probability)

Let's examplify with rats

A probability distribution describes the probability of all potential outcomes

- We kept a 1000 rats in a cage and a number of them died (failure) while others are still alive (success).

$\Rightarrow$ *How can we model this?*

Step 1: describe all potential outcomes

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

```
trials <- 1000
success <- 0:1000
failure <- trials - success
```

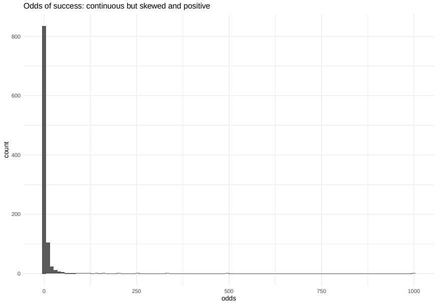
⇒ *We describe all potential outcomes*

Step 2: we calculate the odds

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

- compare successes with failures by dividing one by the other

```
odds <- success/failure
```



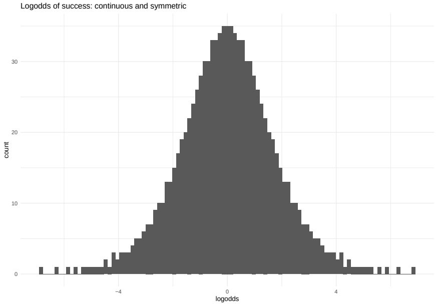
⇒ *A continuous outcome from 0 to + infinity*

Step 3: we log-transform the odds

We logtransform the odds of surviving in a cage in a 1000 cages

- use the logarithmic transformation: natural logarithm (e) of the odds

```
logodds <- log(odds)
```



$\Rightarrow$ *A continuous, bell-shaped outcome from - to + infinity*

Now, let's logtransform the odds

This, we can run regressions on!

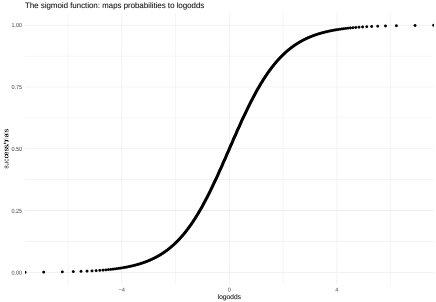
- the outcome variable in logistic regressions is logodds
- ... meaning the regression coefficients are reported on that scale

⇒ *... but they're not easy to understand, so we backtransform when interpreting*

The famous S shape (sigmoid shape)

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

- we can go back and forth between logodds and successes/probabilities
- log-transformation:
  - forces outcome to be between 0 and 1
  - residuals are homoscedastic (constant variance)



⇒ *curve "flattens out" when closing up to the 0 or 1 boundary, so relationship is non-linear*

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 (spread) of the distribution.
- The Bernoulli distribution: $Ber(p)$: when we only have only one trial.

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_i)
- In ML: The z_i must follow a known probability distribution.

$\Rightarrow$ *This what allows us to translate the latent variable to outcomes.*

What happens if I run a linear model on binary outcomes?

- The model risks predicting out of the possible boundaries
  - Predictions are wrong.
  - Regression coefficients are wrong.
  - Standard errors are wrong.
- The relationship between x_i and y_i is constant across all values.

$\Rightarrow$ *This last element has a bearing for the interpretation.*

Example

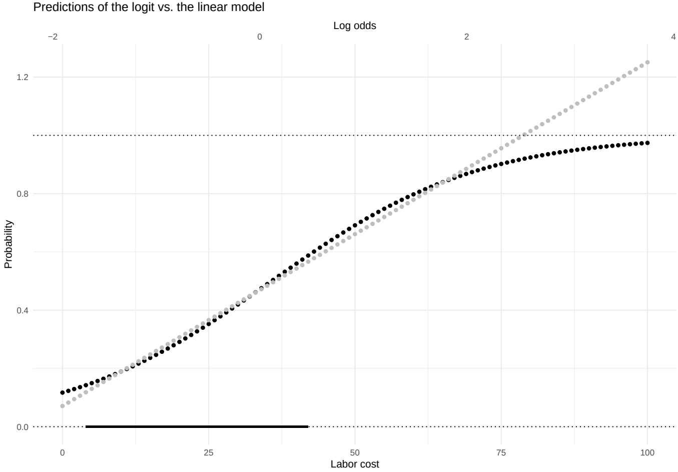
What is the likelihood that MEPs share local assistants, given the cost of employing the?

Table 1: MEP's probability of sharing resources

	Dependent variable:	
	y	
	OLS	*logistic*
	(1)	(2)
LaborCost	0.012^{***}	0.057^{***}
	(0.002)	(0.008)
Constant	0.071^{*}	-2.021^{***}
	(0.041)	(0.224)
Observations	707	707
R^2	0.077	
Adjusted R^2	0.075	
Log Likelihood		-430.848
Akaike Inf. Crit.		865.696
Residual Std. Error	0.460 (df = 705)	
F Statistic	58.479^{***} (df = 1; 705)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Let's back-transform and plot predictions

If we create scenarios for labor cost, we see that at the fringes, the two curves differ.



Interpretation: So. . . what did I find?

Back and forth: Logistic and logit transformation

Back and forth: Logistic and logit transformation

The logit transformation

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

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

$\Rightarrow$ *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.

$$logit^{-1}(logodds) = \frac{exp(logodds)}{1 + exp(logodds)} = \frac{1}{1 + exp(-logodds)} \quad (2)$$

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

My three stages of interpretation

My three stages of interpretation

I go through tree stages of interpretation by first setting two scenarios (or more)

- Marginal effects from regression table
  - Logodds: check direction and significance (in text).
  - Odds ratio (for large coefficients) and percentage change (for smaller coefficients).
- First-difference: predictions with point estimates (in text)
- Predictions: a bunch of scenarios with uncertainty (graphics).

The regression table: marginal effects

I interpret the regression coefficient itself

- Change in logodds: check direction and significance.
- Odds ratio (for large coefficients) and percentage change (for smaller coefficients).

$\Rightarrow$ *A first stab at hypothesis testing.*

The regression table: marginal effects

Now, you try! What statements would you make using the change in logodds, the odds ratio and percentage change? {

Table 2: MEPs' propensity to share local assistants (a binomial logit)

	Dependent variable:
	PoolsLocal
OpenList	-1.124^{***}
	(0.181)
SeatsNatPal.prop	-1.930^{***}
	(0.527)
LaborCost	0.056^{***}
	(0.009)
Constant	-1.094^{***}
	(0.286)
Observations	686
Log Likelihood	-392.832
Akaike Inf. Crit.	793.665
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

}

{

The regression table: marginal effects

Typical statements about marginal effects

- Change in logodds: "MEPs from candidate-centered systems are less likely to share local assistants. Both effects are statistically significant."
- Percentage change (for smaller coefficients; -1.93)."The likelihood that an MEP shares a local assistant with a party colleague is 68% lower when they compete in a candidate-centered system compared to those that compete in party-centered systems."

$\Rightarrow$ *A first stab at hypothesis testing.*

Predicted values

If you believe the model describes reality appropriately, you can learn more about it by interpreting more thoroughly

- Odds ratios are notoriously hard to understand.
- The effect depends on the value of y_i and all the other xs.

$\Rightarrow$ *Interpret the predicted values*

Predicted point estimates (text)

Formulate scenarios using point estimates (in text)

- Take an all-else-equal approach: Let one x change and keep all others constant (on a typical value).
- Find the typical representative of two x values and set the other xs accordingly.

⇒ *Which one you use depends on your objective: A theoretical point, assess effect of intervention on groups...*

Predicted values (graphic)

Formulate scenarios using point estimates and put them on speed

- Predict y values for the entire range of x and plot it.
- Simulate confidence and plot that too.
- You can do this for two scenarios.

$\Rightarrow$ *You get a sense of the actual differences in the data.*

Model assessment: How well is reality described?

Model assessment

Model assessments aim to gauge how well we describe the data (i.e. the y).

- comparison between predicted and observed values (as in OLS).
- mapping outcomes to the recoded, "latent" variable (GLM).

$\Rightarrow$ *You have a few additional "tricks" to the standard OLS assessment.*

Brier score

Describes the "average size" of the residuals.

$$B_b \equiv \frac{1}{n}\Sigma_{i=1}^{n}(\hat{\theta}_i - y_i)^2 \tag{3}$$

$\Rightarrow$ *Lower scores imply better predictions.*

How well do I discriminate?

How well do I discriminate?

The real question for logits is how well do I distinguish 0s from 1s.

- what is the value of my cut point (τ)?

$\Rightarrow$ *Several strategies.*

Table comparison

I can set a single cut point.

- I often use the null-model (i.e. proportion of successes)
  - then recode all probabilities higher than the cut point to 1 and all below to 0:
- How often do I predict correctly?
- on average (proportion of corrects)
- for each value of the outcome (true/false positives and negatives)

$\Rightarrow$ *I can decide how risk-averse I am in my positive predictions*

The ROC curve

The ROC lets the cut values vary and displays how wrong we are on each side (true positive vs. false positive).

- A model with good predictions has a curve tending towards the upper left corner.
- The actual cut value depends on our priorities

$\Rightarrow$ *The graphic is useful in and of itself*

Hosmer-Lemeshow test

Doesn't set the cut point, but bins the data.

- sorts data from low to high probability
- slices it up in g number of groups (e.g. by deciles)

$\Rightarrow$ *performs a* χ^2 *test to assess whether the prediction are significantly different from the observations*

The separation plot

The separation plot shows how the density of observed "successes" increases as our predicted values increase.
$\Rightarrow$ *Another graphic that is useful in and of itself*