Intro to GLM: Binary, Ordered and Multinomial Logistic, and Count Regression Models

Federico Vegetti
Central European University

ECPR Summer School in Methods and Techniques

Goals of the course

By the end of this course you should have learned

- How GLM works in general (and how it is implemented)
- How to analyze several common non-linear dependent variables
- How to interpret results of GLMs
- How to present results in a compellign way

Structure of the course

- Monday: Introduction, binary response variables
- Tuesday: How GLM works in general, Maximum Likelihood Estimation
- Wednesday: Results interpretation and quantities of interest
- Thursday: Categorical and ordered response variables
- Friday: Count variables

General considerations

- Usually our theories are about relationships between concepts
- Concepts are measured, so we test relationships between variables
- Modeling is
    1. Describing a relationship between variables
    2. Describing how our concepts are measured, AKA how the data are generated
- GLM takes into account both aspects

Describing relationships between variables

- Suppose we want to study the relationship between education and income: more educated people have higher-paid jobs
- We measure income as the monthly net salary in Euro
- We measure education as the number of years spent in full-time education
- In our model, the total variation of income consists of:
    1. A **systematic** component: how income varies as a function of education
    2. A **stochastic** component: what is due to other causes, which we can not explain with our data
- A model is a summary of the data in terms of the systematic effect + a summary of the magnitude of the unexplained or random variation

Describing relationships between variables (2)

- A **linear** model is an assumption about the nature of the relationship between income an education
- It describes how much income changes *on average* for a unit increase in education
- It also describes how much of the variation of income is *not* explained by education

$$y_i = X_i\beta + e_i$$

- Where the systematic part is the average of Y given a value of X

$$\mu = E(y|X) = X\beta$$

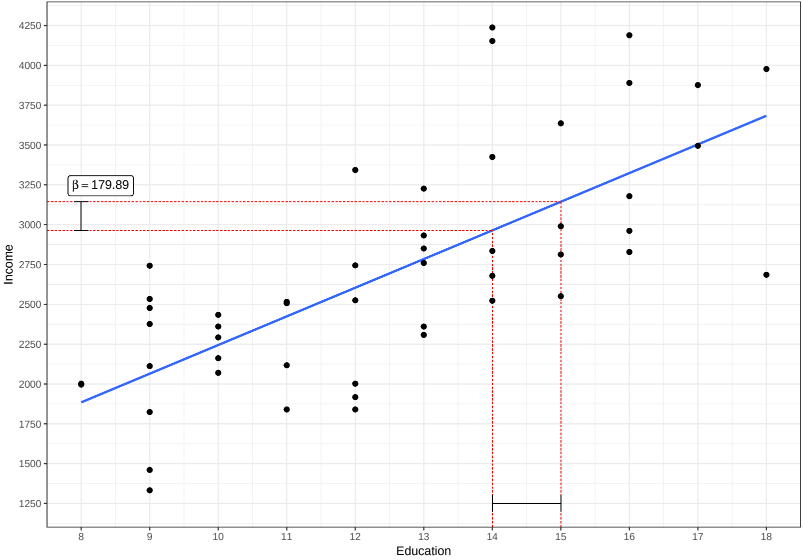
- And the stochastic part is what is left unexplained

$$e_i = y_i - X_i\beta$$

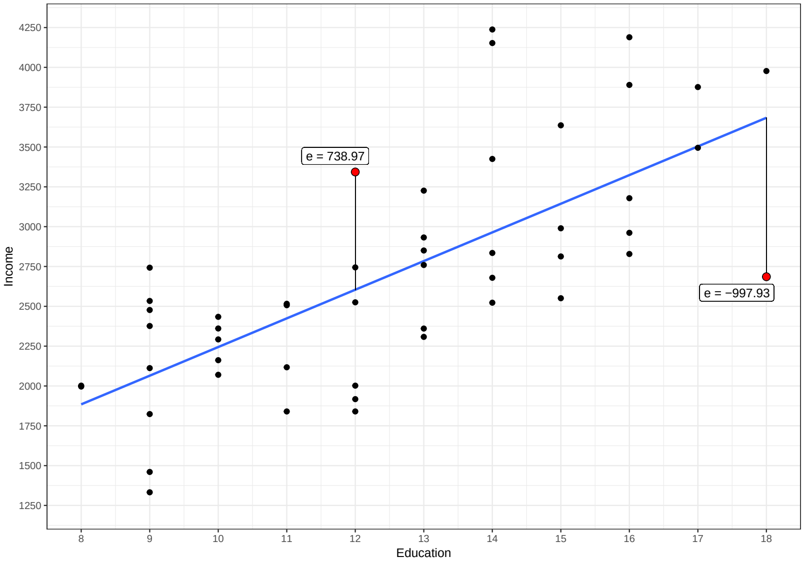
Conceptually

- The systematic component defines the relationship between X and Y, between education and income
  - It looks at the variation of education to explain the variation of income
  - This is what our theories are (usually) about
- The stochastic component defines the distribution of Y
  - It describes the variation of income
  - When we have no predictors (i.e. when we do not know anything about education), *all* the variation of income is stochastic
  - We specify this component by making assumptions about the statistical process that generated the values of income
  - In linear models it is assumed to be "normal"

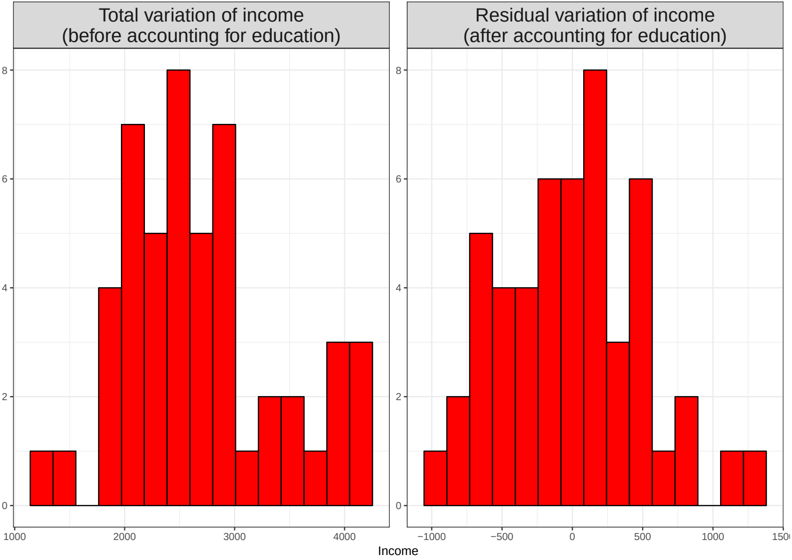
β in practice



e in practice



e in practice (2)



Taking into account how data are generated

- Many social or political event take the form of a yes/no occurrence
  - Did a citizen vote or not?
  - Did a voter choose to vote for the government or for the opposition?
  - Does a person have a job or not?
- What concept do we want to explain here?
- How can we relate other concepts (i.e. independent variables) to it?

The linear probability model

- Sure it is possible to analyze binary responses using linear regression
- This type of model is called **linear probability model**
- Let's consider a voter who has to choose between voting for the *incumbent* party or the *opposition* party

$$y = \begin{cases} 1 & \text{if the incumbent is chosen} \\ 0 & \text{if the incumbent is not chosen} \end{cases}$$

- We can model y as a linear function of people's economic situation compared to the year before
- The more their finances have improved (the higher the value of X) the more likely they will vote for the government

$$y_i = X_i\beta + e_i$$

The linear probability model (2)

- The linear model implies that

$$E(y) = X\beta$$

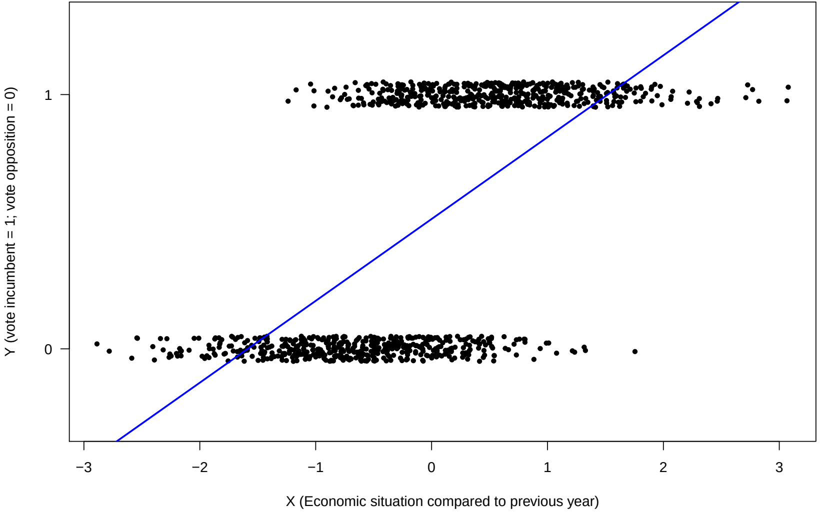
- *E(y)* is the mean of y, which is just the share of y*=1* in our data
- This is interpreted as a probability

$$E(y) = P(y = 1) = \pi$$

- I.e. the linear probability model predicts the mean of y, which is the probability that y has value 1
- It is interpreted in the same way as with linear regression: for 1 point increase in X, β tells how much the probability that y*=1* (that is π) increases

LPM in practice

Example: $Y = 0.51 + 0.32X$



Problems with the LPM

- Besides the violation of normality and homoskedasticity assumptions (which can affect the validity of our results) there are two more immediate concerns:
  1. The LPM makes out-of-bounds predictions
  2. The linear functional form might apply badly to a concept like probability
- The first point is straightforward: what's the predicted value of Y when $X = -2$?
- The second point is trickier
  - The linear functional form implies that π changes at a constant rate, regardless the starting point of the predictor
  - However, this is hardly the case

On probability change

- Example: Bill is choosing whether to buy a product that costs 5€
- One factor influencing the decision is Bill's wealth (X)
- We give him 1€, AKA we increase X of 1 unit
- How much does the probability that Bill buys the product change?

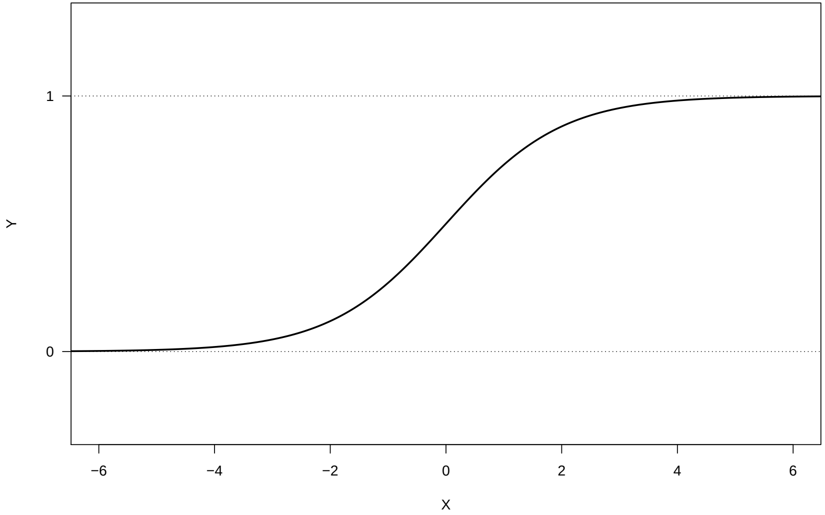
On probability change (2)

- Bill has 0€:
  - Not a great improvement. Bill is still short of 4€, so the probability that he buys the product won't change much
- Bill is millionaire:
  - If he didn't buy the product yet, it's not because of money. Probably he doesn't need it, or he doesn't like it. Again, the change in probability as X increases 1 point will be small
- Bill has 4€
  - Now things are different. By giving Bill 1€, we change his state from not being able to afford the product to being able to do so. Increasing X of 1 unit at this point could have a huge effect

The functional form

- The functional form describes how X relates to Y
- When we model a probability change, we are in fact modeling a discrete event
- This implies that all the possible change of Y can be realized only in one single "step" from 0 to 1
- For this relationship, a **sigmoid** functional form is more appropriate
  - For very low values of X, any increase will have a relatively little impact
  - As we move along the range of X, the effect of one unit increase becomes larger and larger
  - However, passed a certain point, the effect of one unit increase in X becomes smaller again
- To specify the correct functional form is a fundamental step in statistical modeling

Sigmoid relationship



Modeling probabilities with GLM

- The most common ways to model binary outcomes rely on this assumption
- How can we work this out? With GLM
- We need to **transform** the probability of Y (i.e. the mean of Y) in a way such that it can be related to X linerarly
- We do this using a mathematical function called **link function**
- The link function transforms a probability into a quantity called **linear predictor**
- The linear predictor is the systematic component of the model, and can be modeled in the same way as in "simple" linear models

GLM in a nutshell

At the most general level, GLM consists of 3 steps

1. Specify the distribution of the dependent variable
   - This is our assumption about how the data are generated
   - This is the stochastic component of the model
2. Specify the link function
   - We “linearize” the mean of Y by transforming it into the linear predictor
   - It always has an inverse function called **response function**
3. Specify how the linear predictor relates to the independent variables
   - This is done in the same way as with linear regression
   - This is the systematic component of the model

Logit and Probit models

- To model probabilities of binary events, we need a function that maps our linear predictor to a cumulative distribution function
- Two common functions are at the basis of the **logit** and the **probit** models
- The two models work exactly in the same way, except they use a different link function
- Let's consider the linear predictor

$$\eta = X\beta$$

- To be mapped to the probability π with a response function *h()*:

$$\pi = h(\eta) = h(X\beta)$$

Logit models

- We need to find a response function that turns a linear unbounded distribution into a distribution that:
  - Is bounded between 0 and 1
  - Relates to X with a sigmoid functional form
- Logit models use the standard logistic cumulative distribution function:

$$\pi = \frac{exp(\eta)}{1 + exp(\eta)} = \frac{exp(X\beta)}{1 + exp(X\beta)}$$

- And the link function is called **logit function**:

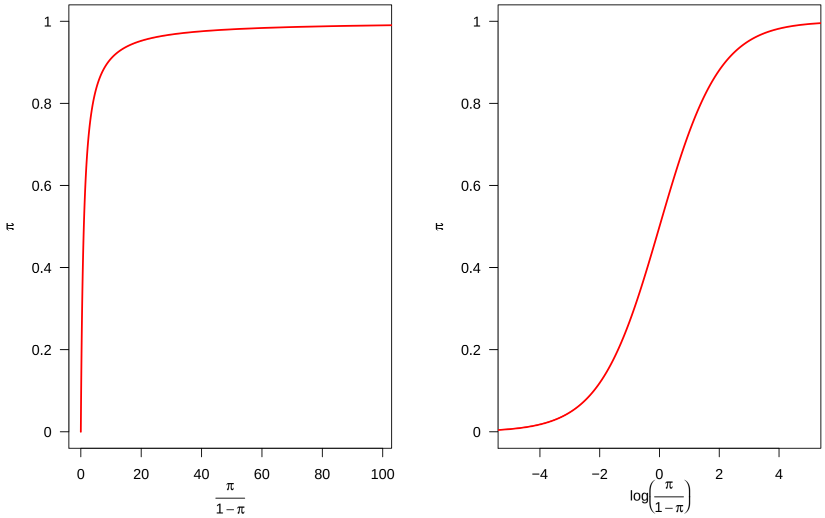
$$\eta = X\beta = log\left(\frac{\pi}{1 - \pi}\right)$$

- The part $\left(\frac{\pi}{1-\pi}\right)$ is called "odds", and refers to the probability to observe an event versus its complement

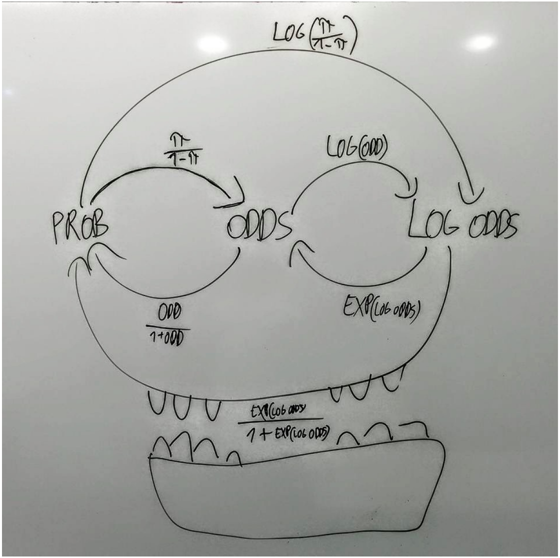
Probabilities, odds, and log odds

Probability π	Odds $\frac{\pi}{1-\pi}$	Logits $log\left(\frac{\pi}{1-\pi}\right)$
0.01	$1/99 = 0.0101$	-4.60
0.05	$5/95 = 0.0526$	-2.94
0.10	$1/9 = 0.1111$	-2.20
0.30	$3/7 = 0.4286$	-0.85
0.50	$5/5 = 1$	0.00
0.70	$7/3 = 2.3333$	0.85
0.90	$9/1 = 9$	2.20
0.95	$95/5 = 19$	2.94
0.99	$99/1 = 99$	4.60

Probabilities, odds, and log odds (2)



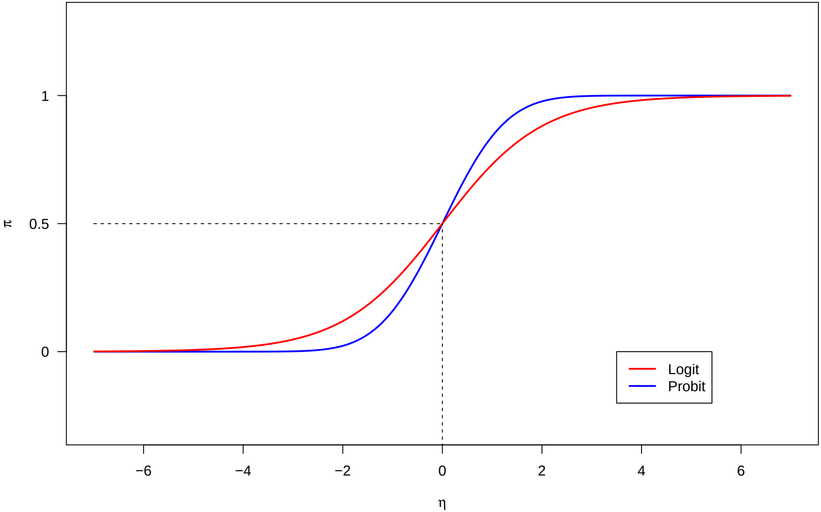
Probabilities, odds, and log odds (3)



Probit models

- In probit models, the response function *h()* is the standard normal CDF:
$$\pi = \Phi(\eta) = \Phi(X\beta)$$
- And the link function *g()* is the inverse:
$$\pi = \Phi^{-1}(\eta) = \Phi^{-1}(X\beta)$$
- However, the inverse function Φ^{-1} has no easy analytic solution, so it is found numerically

Logit vs. Probit functions



Logit and Probit models

- Note from the figure that both functions are nearly linear for the most of their range
  - In fact the linear probability model leads to similar results, except for extreme values of Y
- Logit and probit models produce identical predicted values, but different coefficients
- Models using the logit link function are more common than probit models
- This is also a matter of ease of interpretation:
- Essentially, logit models are **linear models for log-odds**

A latent variable interpretation

- Binary response variables can be regarded more directly as a measurement problem
- We can think of a continuous unobservable construct $y*$, e.g. the propensity to turnout at the next election
- We can't observe $y*$, we can only observe its manifest variable y in two states, e.g. whether a persone says s/he will vote at the next election or not
- In fact, a voter might be barely convinced to turn out, while another might be enthusiastic about the election
- However, all we see is the discrete choice whether they will vote (1) or not (0)

A latent variable interpretation (2)

- $y*$ is linked to y by the measurement equation:

$$y_i = \begin{cases} 0 & \text{when } y*_i \leq 0 \\ 1 & \text{when } y*_i > 0 \end{cases}$$

- The value *0* is an arbitrary threshold on $y*$: when it is passed, y switches from *0* to *1*
- In this context we model:

$$y*_i = X_i\beta + e_i$$

- And the probability that $y_i = 1$ is:

$$P(y*_i > 0) = P(X_i\beta + e_i > 0)$$

A latent variable interpretation (3)

- Since $y*$ is not observed, we can't estimate its variance: we need to fix it at a given value
- Different assumptions about the variance of e lead to different model specifications:
  - If $Var(y*) = \pi^2/3$, $y*$ follows a standard logistic distribution
  - If $Var(y*) = 1$, $y*$ follows a standard normal distribution
- Depending on which distribution of e we assume, solving the equation in the previous slide produces formulations that are equivalent to the logit or the probit model
- This approach requires more theorization – i.e. we need to find a convincing definition of the latent variable
- However, in practice it produces identical results

In sum

- Binary responses can not be related to our predictors linearly
- To model them, we need to transform their distribution in a way that can be treated as in a linear model
- GLM requires us to:
  - Make an assumption about the distribution of y
  - Find a link function to make the distribution of y linear
  - Model the transformed linear predictor

What we will see tomorrow

- How GLM for binary responses works with individual data
- How parameters in GLM are estimated