

Applied Statistics

Lecture 5

Prof.ssa Chiara Seghieri, Dott.ssa Costanza Tortù

Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS
Scuola Superiore Sant'Anna, Pisa

c.seghieri@santannapisa.it

c.tortu@santannapisa.it

Outline

1. Motivation and Intuition
2. Why the standard regression model is not appropriate with binary outcomes
3. Introduction to GLMs
4. The LOGIT Model
5. Interpretation of coefficients
6. Inference on the coefficients
7. Goodness of fit
8. Other common GLMs: the Probit and Poisson Model

1) Motivation and intuition

Introduction

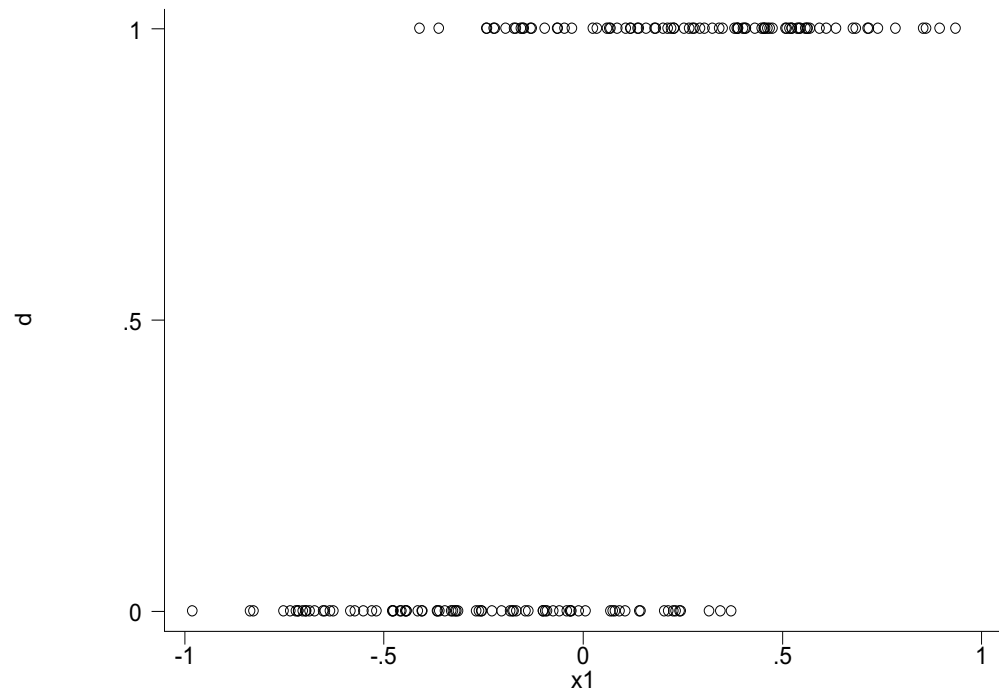
Sometimes the dependent variable Y is can take two mutually exclusive values:

- Y = get into college (success), or not (failure); X = gender
- Y = person smokes, or not; X = income
- Y = mortgage application is accepted, or not; X = income, house characteristics, marital status, race

The goal is to describe the way in which the probability of Y (i.e. get into college) varies by X .

Y is a random bernoulli variable taking two values: 0,1.

We know that for Y : $E(Y)=p$ and $\text{var}(Y)=p(1-p)$



The aim is to model this relation:

$$p_i = P(y_i = j) = F(x_i' \beta)$$

y_i is the binary dependent variable, x_i is the vector of the k regressors, β the vector of k parameters and p_i is the probability that the observation i chooses the alternative j (0,1). $i=1, \dots, N$

2) Why a linear model is not appropriate for binary outcomes

Linear Probability Model

The most obvious idea is to let the probability of success be a linear function of x .

$$E(y_i|x) = P(y_i = 1|x = x_i) = \alpha + \beta x \quad i=1,2,\dots,n$$

the variance of Y is not constant!

The right hand takes values in the range $(-\infty, +\infty)$

The left hand ranges between 0 and 1

$$\text{var}(y_i) = E(y_i) * (1 - E(y_i))$$

Linear Probability Model

We can estimate such a model by using OLS.

- However, we don't get good results:

Probabilities fall between 0 and 1, but linear functions take values over the entire real line.

During the fitting process, $p(x)$ falls outside the (0,1) range for some x values. The model can be valid over a restricted range of x values.

Problems with LPM

Moreover, LPM assumes that probabilities increase linearly with the explanatory variables

- Each unit increase in an X has the same effect on the probability of Y occurring regardless of the level of the X .
- Moreover, in many situations we empirically see “diminishing returns”
 - changing p by the same amount requires a bigger change in x when p is already large (or small) than when p is close to $1/2$. Linear models can’t do this.
- As Aldrich and Nelson (1984) note, we want a model that “approaches zero at slower and slower rates as X gets small and approaches one at slower and slower rates as X gets very large.”

EX: Let $p(x)$ denote the probability of buying a new house when annual family income is x . An increase of \$50,000 in annual income would have less effect when x is \$1,000,000 (for which $p(x)$ is near 1) than when x is \$50,000.

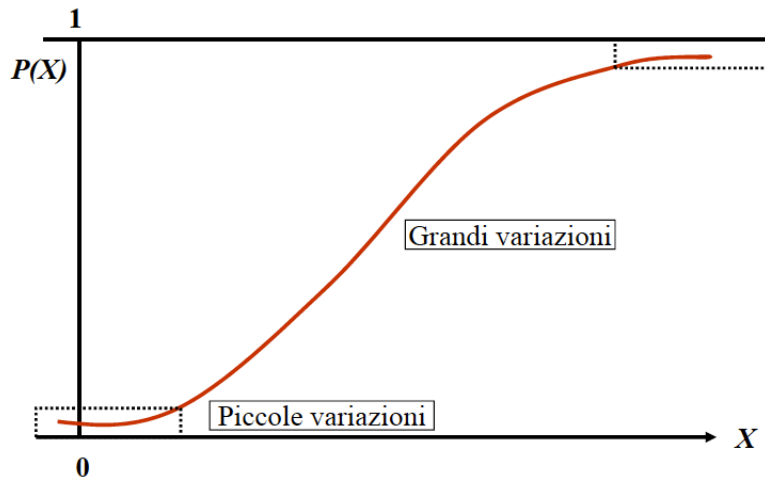
We therefore need to find a functional form for $F(x_i'\beta)$ for which:

$$\begin{cases} \lim_{x_i'\beta \rightarrow -\infty} F(x_i'\beta) = 0 \\ \lim_{x_i'\beta \rightarrow +\infty} F(x_i'\beta) = 1 \end{cases}$$

► Any distribution function (CDF) of a continuous random variable is suitable for this objective. Since $F(x_i'\beta): \mathbb{R} \rightarrow [0,1]$

The **logistic** distribution and **standard normal** distribution are two candidates. The logistic gives rise to the **logit** model; the standard normal gives rise to the **probit** model

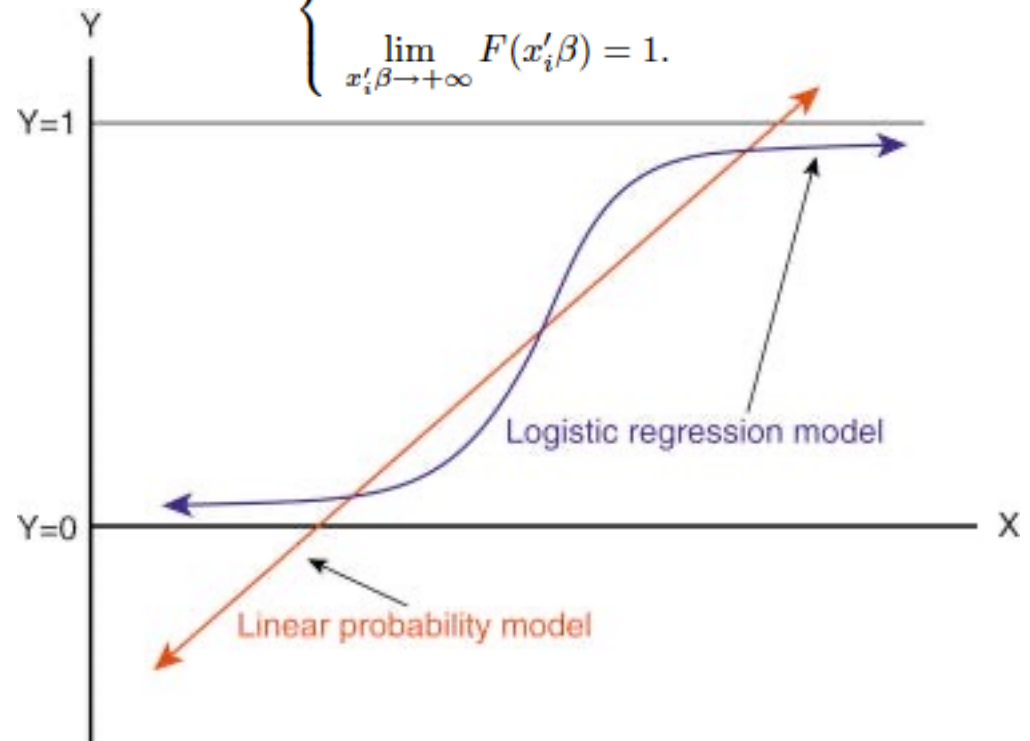
The logit model



β indicates if $[P(y = 1)]$ increase ($\beta > 0$) or decrease ($\beta < 0$) for an increase x

$$p_i = \text{pr}(y_i = 1) = \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}$$

$$\begin{cases} \lim_{x'_i \beta \rightarrow -\infty} F(x'_i \beta) = 0 \\ \lim_{x'_i \beta \rightarrow +\infty} F(x'_i \beta) = 1. \end{cases}$$



3) Introduction to GLMs

Motivation

In the standard linear models we look for a linear predictor for expected value of the response variable.

BUT - as we have hinted -linear models have limitations

- Hypotheses are not always valid
- Necessity to study non – linear relationships
- Problems with heteroskedastic data

Solution: jump to the **Generalized Linear Models** where the linear predictor models **a function of** the Expected Value of the Y

Linear model

$$E[y|x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$



Generalized Linear model (1972)

$$g(E[y|x]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

What characterizes a GLM

The three elements that fully characterize a GLM

- The statistical distribution of the responses Y is

- The link function $g(\cdot)$

Idea: when the response has a nonlinear relationships with predictors, a transformation of the response is expressed as a linear regression

- The linear predictor

$$\eta = g[E(y)] = \underbrace{\beta^T x}_{\text{Linear predictor}}$$

Diagram illustrating the components of the GLM equation $\eta = g[E(y)] = \beta^T x$:

- η (green) is the linear predictor.
- g (red) is the Link function.
- $E(y)$ (purple) is the Expected value of the response variable.
- $\beta^T x$ (black) is the Linear predictor.

Note: LMs can be regarded as a particular GLM

- General GLM

$$Y = f(X_1, X_2, X_3) + \varepsilon$$

- Linear Models

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$



A linear model is a generalized linear model where the link function is the **identity function**

What is a GLM

GLIMs are a family of models that:

- Extend linear regression to a broader family of outcome variables
- Enables researchers to use the linear modelling framework to variables that are not Normally distributed.

<i>Family</i>	<i>Default "Link"</i>	<i>Range of y_i</i>
gaussian	identity	$(-\infty, +\infty)$
binomial	logit	$\frac{0, 1, \dots, n_i}{n_i}$
poisson	log	$0, 1, 2, \dots$
Gamma	inverse	$(0, \infty)$
inverse gaussian	$1/\mu^2$	$(0, \infty)$

4) The LOGIT Model

The logit model

$$Y \sim \text{Bin}(n, p)$$

Probability form

$$P(Y = 1) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

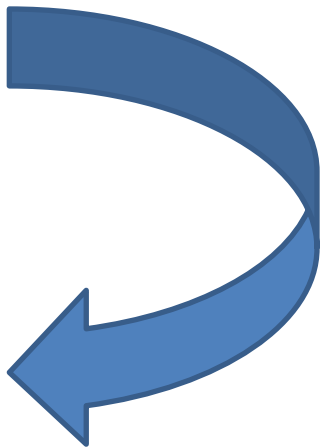
we linearize the expression by applying the natural log to the two members of the equation (Logit form):

$$\log \frac{p(Y = 1)}{1 - p(Y = 1)} = \alpha + \beta x$$

Logistic regression

$P(Y=0)$

N.B.: This is natural log (aka “ln”)




Probabilities, odds, and logits

The logit model has three equivalent forms:

Probabilities: $P(Y=1) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$

Logits (Log-Odds): $\ln \frac{P(y=1)}{P(y=0)} = \alpha + \beta x$

$\text{logit}[P(y=1)] = \alpha + \beta x$


= 1 - P(y=1)

Odds: $\frac{P(y=1)}{P(y=0)} = e^{\alpha + \beta x}$

Odds, Odds Ratios

- **The odds** of “success” is the ratio: $\omega = \frac{p}{1-p}$
- consider two groups with success probabilities:
 p_1 and p_2
- **The odds ratio** (OR) is a measure of the odds of success in group 1 *relative* to group 2

$$\theta = \frac{\omega_1}{\omega_2} = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)}$$

Odds for Independent Variable Groups

- We can compute the **odds of receiving a death penalty** by race:

	Blacks	Nonblacks	Total
Death sentence	28	22	50
Life imprisonment	45	52	97
Total	73	74	147

- The odds of receiving a death sentence if the defendant was Black = $p/1-p = (28/73)/(1-(28/73)) = 0.6222$
- The odds of receiving a death sentence if the defendant was not Black = 0.4231

The Odds Ratio Measures the Effect

- The impact of being black on receiving a death penalty is measured by the odds ratio which equals:
 - = the odds if black / the odds if not black
 - = $0.6222 / 0.4231 = 1.47$
- Which can be interpreted as:
 - Blacks are 1.47 times more likely to receive a death sentence as non blacks
 - The risk of receiving a death sentence are 1.47 times greater for blacks than non blacks
 - The odds of a death sentence for blacks are 47% higher than the odds of a death sentence for non blacks. ($1.47 - 1.00$)
 - A one unit change in the independent variable race (nonblack to black) increases the odds of receiving a death penalty by a factor of 1.47.

Maximum likelihood estimates

- Logit and Probit models are nonlinear in the coefficients β therefore these models can't be estimated by the standard OLS (you could use non linear OLS but not efficient!)
- MLE is an alternative to OLS. It consists of finding the parameters values which is the most consistent with the data we have.
- In Statistics, the likelihood is defined as the joint probability to observe a given sample, given the parameters involved in the generating function.
- One way to distinguish between OLS and MLE is as follows:

OLS adapts the model to the data you have: you only have one model derived from your data. MLE instead supposes there is an infinity of models and chooses the model most likely to explain your data.

The method of maximum likelihood yields values for the unknown **parameters** that **maximize the probability of obtaining the observed set of data**

5) Interpretation of coefficients

Parameter interpretation

$$\log \frac{p(Y = 1)}{1 - p(Y = 1)} = \alpha + \beta x$$

Logistic slope coefficients can be interpreted as the **effect of a unit of change in the X variable on the logits** with the other variables in the model held constant. That is how a one unit change in X effects the log of the odds when the other variables in the model held constant.

Parameter interpretation

- Exponentiating both sides of the logit link function we get the following:

$$\left(\frac{p_i}{1-p_i} \right) = \text{odds} = \exp(\beta_0 + \beta_1 X_1) = e^{\beta_0} e^{\beta_1 X_1}$$

- The odds increase **multiplicatively** by e^{β_1} for every 1-unit increase in x , x continuous.
- The odds at $X = x+1$ are e^{β} times the odds at $X = x$, furthermore, $(e^{\beta} - 1) * 100$ gives the percent increase in the odds of a success for each 1-unit increase in x if estimate of $\beta > 0$. $(1 - e^{\beta}) * 100$ estimate of $\beta < 0$

Parameter interpretation

for x (*i.e. income in thousands of dollars-continuous*) and Y =buying a house or not. The estimate of β is equal to 0.012. If income is increased by \$10, this increases the odds of buying a house by about 13%

$$(e^{10 \times 0.012} - 1) \times 100\% = 12.75\%$$

- if estimate of $\beta > 0$ then percent increase in odds for a unit change in x is

$$(e^{\hat{\beta}} - 1) \times 100\%$$

- if estimate of $\beta < 0$ then percent decrease in odds for a unit change in x is

$$(1 - e^{\hat{\beta}}) \times 100\%$$

Parameter interpretation

Example: for x (*dummy variable coded 1 for female, 0 male*), y satisfaction. The odds ratio is

$$\theta = \frac{p_f / (1 - p_f)}{p_m / (1 - p_m)} = \frac{\omega_f}{\omega_m} = \exp(\hat{\beta}_2) = \exp(0.67) = 1.95$$

- holding the other variable constants, women's odds of buying a house is nearly twice those of men.

Marginal effects

Marginal effects: change in the probability of the event occurring (e.g., the probability of success) due to **a one-unit change in an independent variable**, while holding all other variables constant.

How to compute marginal effects?

1. Fit the logit model model to your data.
2. Once you have estimated the model parameters (the β coefficients), you can calculate the marginal effects for each independent variable. The marginal effect for an independent variable X_j is calculated as:

3. Marginal Effect (X_j) = $\beta_j * p * (1 - p)$

Where:

- β_j is the coefficient of the independent variable X_j .
- p is the predicted probability of the event happening, given the values of all the independent variables. You can obtain this by plugging the values of your predictors into the logistic regression equation.

6) Inference on the coefficients

The model equation

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6590	0.2880	33.1855	<.0001
AGE	1	0.0285	0.00838	11.5255	0.0007

$$P(y = 1) = \frac{\exp(-1.659 + .0285x_1)}{1 + \exp(-1.659 + .0285x_1)}$$

$$\ln \left(\frac{\pi}{1 - \pi} \right) = -1.659 + .0285x_1$$

$$\text{logit}(y) = -1.659 + .0285x_1$$

Inference: The Coefficients.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6590	0.2880	33.1855	<.0001
AGE	1	0.0285	0.00838	11.5255	0.0007

Instead of a t -test for the significance of a coefficient (like in linear regression), we have a Wald Chi-Squared test.

Inference: The Coefficients.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6590	0.2880	33.1855	<.0001
AGE	1	0.0285	0.00838	11.5255	0.0007

Furthermore: $(\exp(.0285)-1)*100\% = 2.89\%$.

We can state that the odds of being in poor health increase by 2.89% with each additional year in age.

Ex: Y=1 not winner (of election)

logit nowin leader age scandal

```
Logit estimates                                Number of obs   =       5036
                                                LR chi2(3)      =       265.97
                                                Prob > chi2     =       0.0000
Log likelihood = -1214.2961                    Pseudo R2      =       0.0987
```

nowin	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
leader	-1.029759	.516245	-1.99	0.046	-2.04158	-.0179374
age	-.0420528	.0035401	-11.88	0.000	-.0489913	-.0351143
scandal	2.839299	.3194128	8.89	0.000	2.213261	3.465337
_cons	-1.487179	.0859136	-17.31	0.000	-1.655566	-1.318791

A positive coefficient: the log-odds of not winning are decreasing as a function of being in a leadership position and with increasing of age and increase as a function of being involved in a scandal.

logistic nowin leader age scandal

```
Logit estimates                                Number of obs   =       5036
                                                LR chi2(3)      =       265.97
                                                Prob > chi2     =       0.0000
Log likelihood = -1214.2961                    Pseudo R2      =       0.0987
```

Nowin	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
leader	.357093	.1843475	-1.99	0.046	.1298234	.9822225
age	.9588192	.0033943	-11.88	0.000	.9521895	.965495
scandal	17.10377	5.463164	8.89	0.000	9.145496	31.98723

Scandal: the odds of not winning of candidate who is involved in a scandal are about 17 times higher than a candidate who is not involved.

Age: $(1 - (.9588)) * 100 = 4\%$ decrease in the odds of not winning for a unit increase in age

7) Goodness of fit

Goodness of Fit Measures

- In ML estimations, there is no such measure as the R^2
- However, the log likelihood measure can be used to assess the goodness of fit.
 - **NOTE** Given the number of observations, the better the fit, the higher the LL measures
- The philosophy is to compare two models looking at their LL values. One is meant to be the constrained model, the other one is the unconstrained model.

Goodness of Fit Measures

- A model is said to be **constrained** when the parameters associated with some variable are set to zero.
- A model is said to be **unconstrained** when the parameters associated with some variable are allowed to be different from zero.
- For example, we can compare two models, one with no explanatory variables, one with all our explanatory variables. The one with no explanatory variables implicitly assume that all parameters are equal to zero. Hence it is the constrained model because we (implicitly) constrain the parameters to be null.

The likelihood ratio test (LR test)

- The most used measure of goodness of fit in ML estimations is the **likelihood ratio**. The likelihood ratio is the difference between the unconstrained model and the constrained model. This difference is distributed χ^2 .
- If the difference in the LL values is (no) important, it is because the set of explanatory variables brings in (un)significant information. The null hypothesis H_0 is that the model brings no significant information as follows:

$$LR = 2 \left[\ln L_{\text{unc}} - \ln L_c \right]$$

- High LR values will lead the observer to reject hypothesis H_0 and accept the alternative hypothesis H_a that the set of explanatory variables does significantly explain the outcome.

Other usage of the LR test

- The LR test can also be generalized to compare any two models, the unconstrained one being *nested* in the constrained one.
- Any variable which is added to a model can be tested for its explanatory power as follows :
 - `logit [model constraint]`
 - `est store [name1]`
 - `logit [model non constraint]`
 - `est store [name2]`
 - `lrtest name2 name1`

The McFadden Pseudo R²

- We also use the McFadden Pseudo R² (1973). Its interpretation is analogous to the OLS R². However it remains generally low.
- pseudo-R² also compares The likelihood ratio is the difference between the unconstrained model and the constrained model and is comprised between 0 and 1.

$$\text{Pseudo } R_{\text{MF}}^2 = \frac{[\ln L_c - \ln L_{\text{unc}}]}{\ln L_{\text{unc}}} = 1 - \frac{\ln L_{\text{unc}}}{\ln L_c}$$

Model Fit Statistics

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	238.329	228.316
SC	241.607	234.873
-2 Log L	236.329	224.316

Both AIC & SC are deviants of the -2 Log L that penalize for model complexity (the number of predictor variables).

Model Fit Statistics

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	238.329	228.316
SC	241.607	234.873
-2 Log L	236.329	224.316

AIC Akaike Information Criterion. Used to compare non-nested models. Smaller is better. AIC is only meaningful in relation to another model's AIC value.

Model Fit Statistics

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	238.329	228.316
SC	241.607	234.873
-2 Log L	236.329	224.316

Choose either AIC or SC (not both) and use the values under the heading 'Intercept and Covariates' to compare to competing models.

Checking assumptions

0. Independent data points

(no tests for that, just think about your data)

Problem: likelihood function is wrong otherwise + confidence intervals too small

1. Influential data points

2. No multi-collinearity

3. All relevant variables included

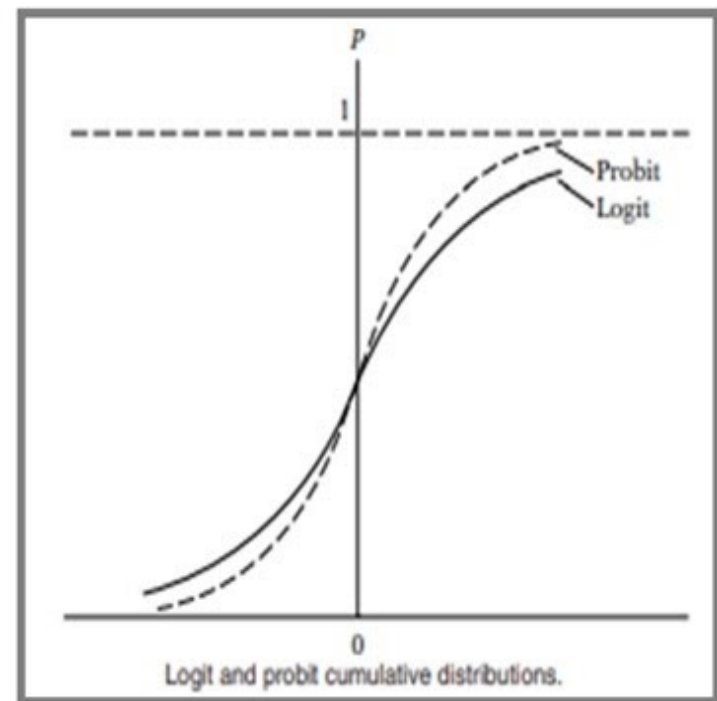
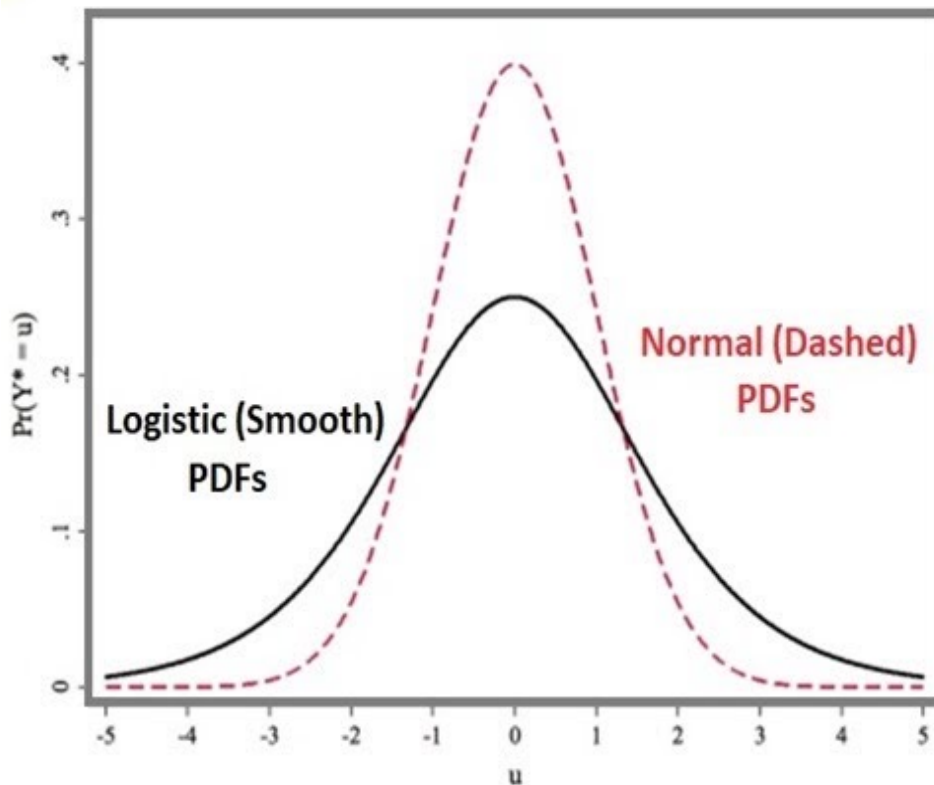
8) Other GLMs: the Probit and Poisson model

1 – Probit Model

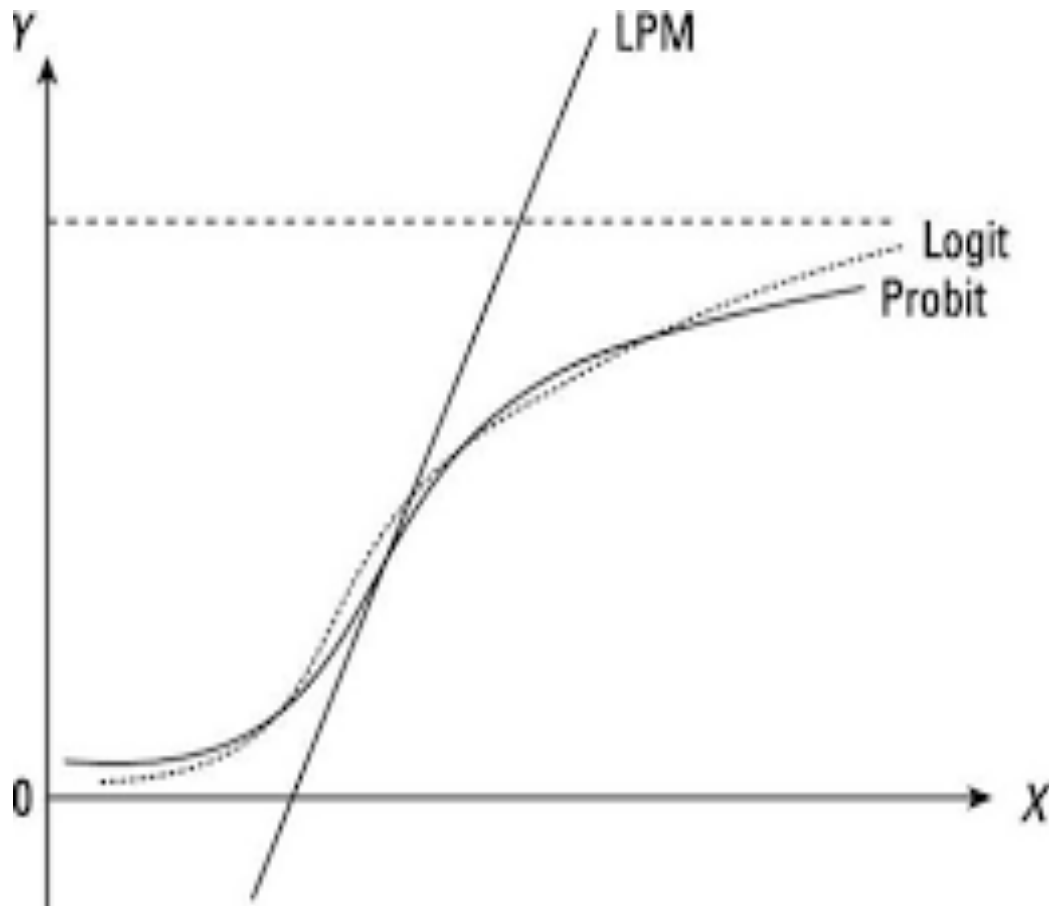
An alternative model for binary outcomes:

Logit model: uses the CDF of the **logistic distribution**

Probit model: uses the CDF of the **normal distribution** --- > **z-scores**



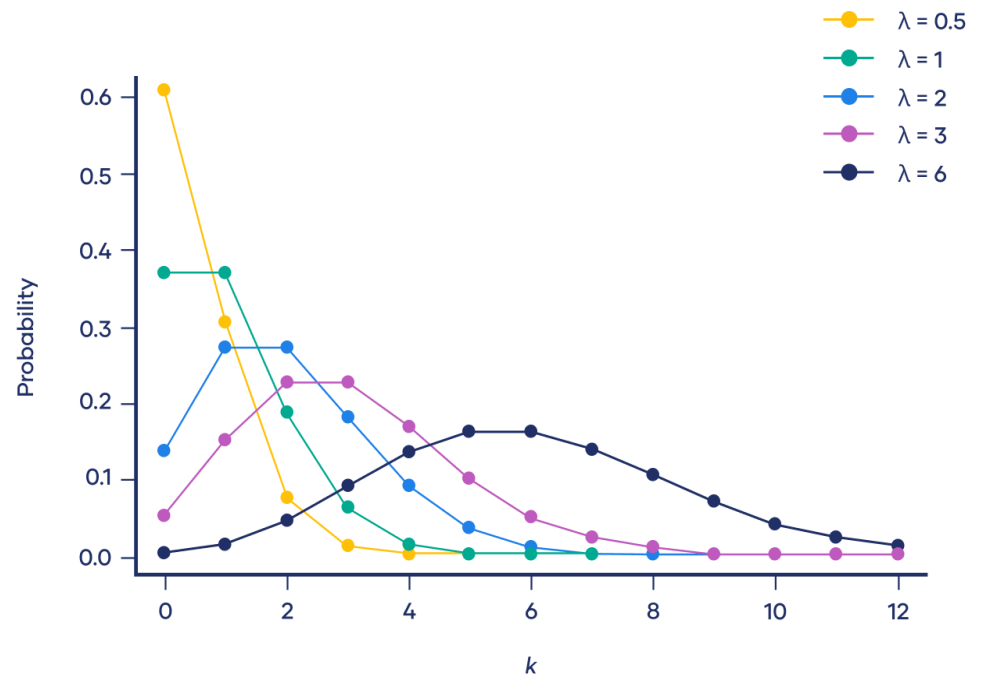
Probit Model vs Logit Model vs LPM



- Both Logit and Probit are interpreted in terms of OR.
- In the **Logit** model the interpretation focuses on **probabilities**, while in the **Probit** model coefficients are interpreted in terms of their effect on the change of **z-scores**.
- The logit model has heavier tails
- As a consequence, the logit model is more robust with respect to outliers.

2 – Poisson Model

- Count response variable (frequencies) in a fixed period of time, with a Poisson distribution
- **Poisson distribution**: probability of $0, 1, 2, \dots$ events; the mean of the distribution is equal to the variance
- The probability of experiencing more than one event is negligible

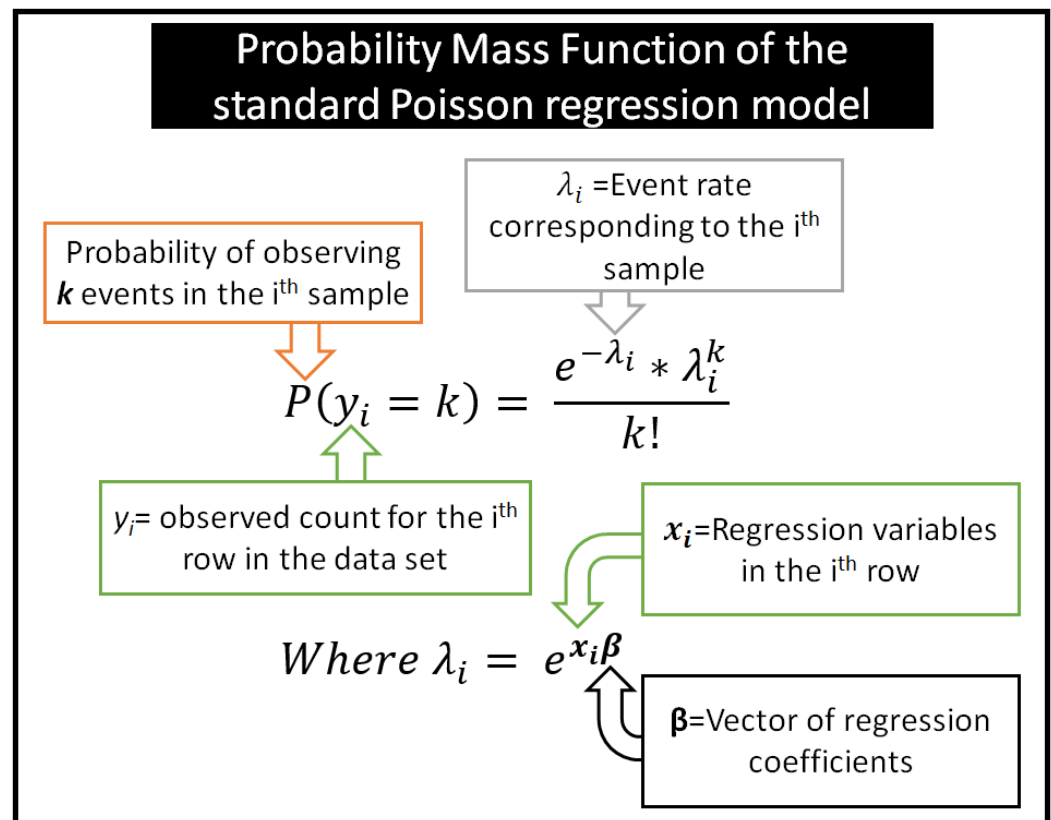


Link function Linear predictor

$$\ln \lambda_i = b_0 + b_1 x_i$$

$$y_i \sim \text{Poisson}(\lambda_i)$$

Probability distribution



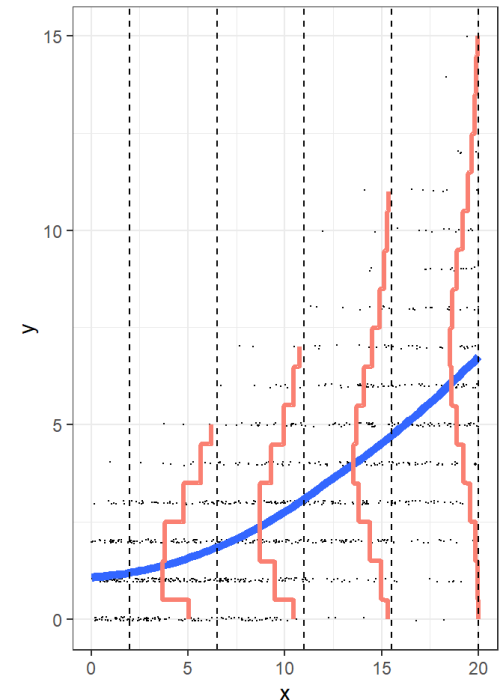
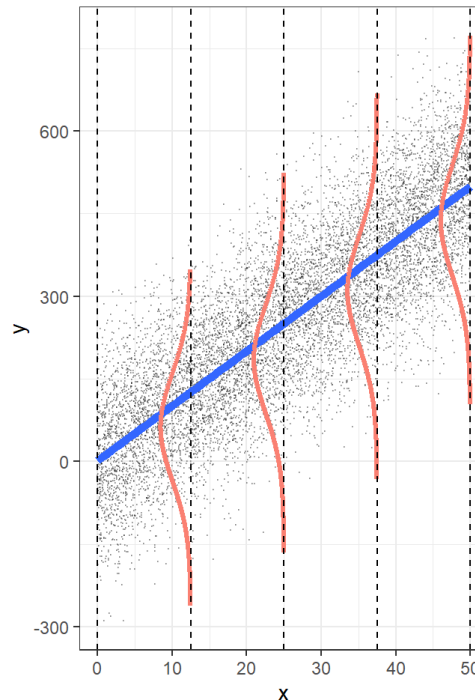
Note: the Poisson distribution can be characterized by **overdispersion** (as the mean increases the variance also increases)

With the Poisson distribution, we are implicitly introducing a source of **heteroskedasticity**

Check standardized residuals

Ideal characteristics of standardized residuals

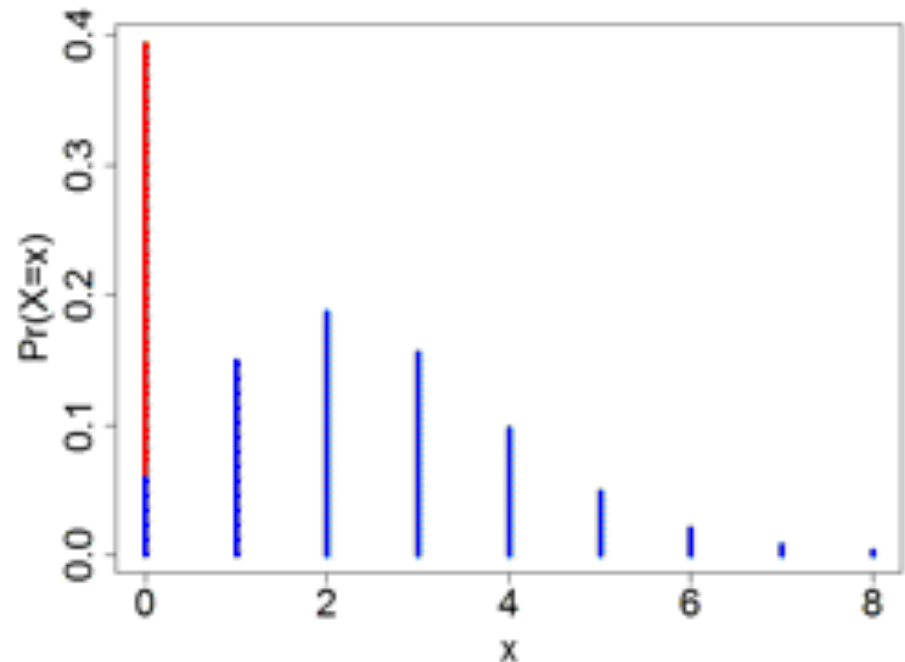
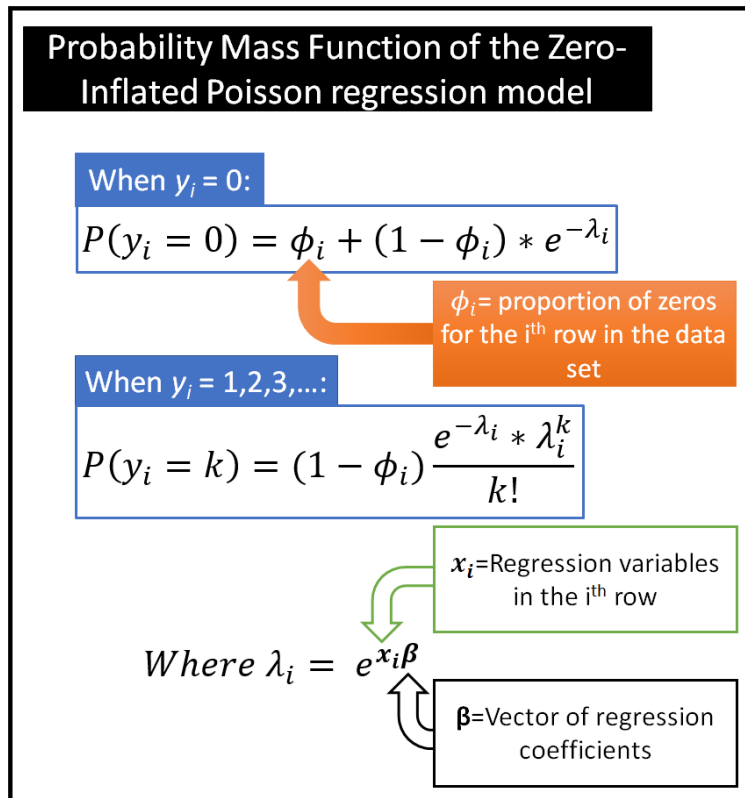
- they should be small
- as N approaches infinity, they should be distributed as a Standard Normal
- their distribution should not reveal specific patterns (wrt their relationship with the y_i)
- their values should be around 0



Issues related to the Poisson regression

A) Issues with count data

- A1) Zero Inflated Distribution: distribution of the Y_i s characterized by a peak on the 0
 - Zero Inflated Poisson (ZIP): mixture distribution that separately models the zeros

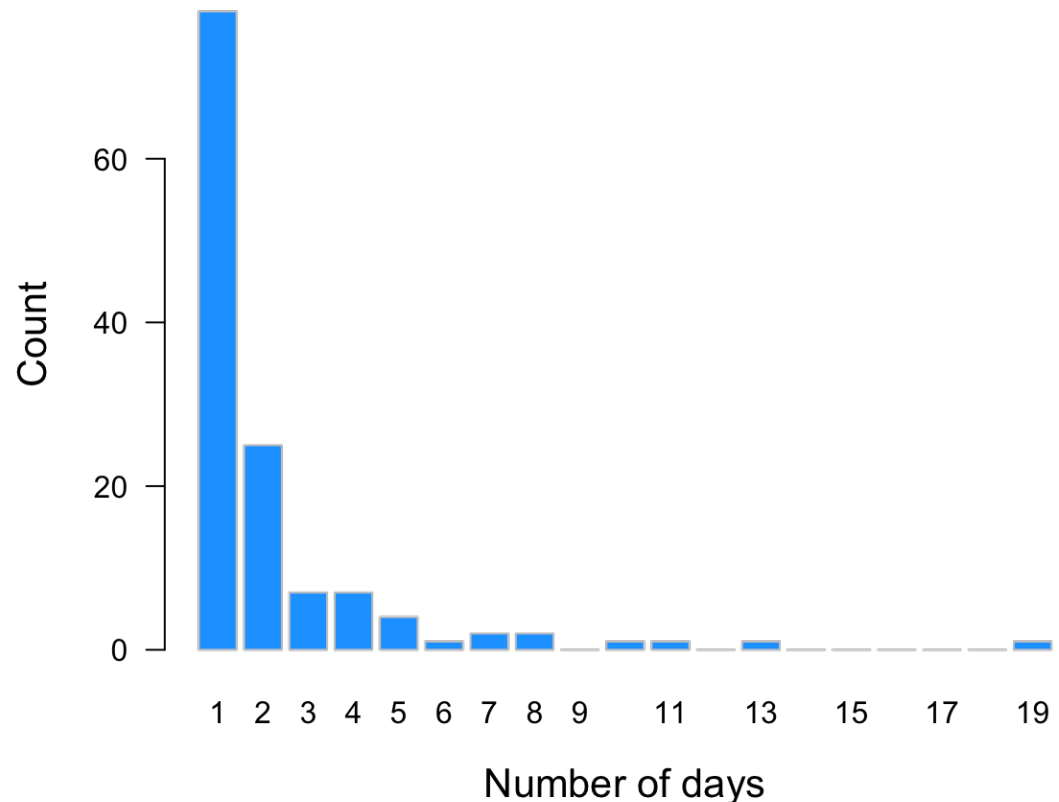


Issues related to the Poisson regression

A) Issues with count data

- A2) Truncated Poisson: The experimental designs suggest that there cannot be 0s, by construction.

→ Work with a slightly different probability function and model $P(Y_i=y | Y_i > 0)$



Issues related to the Poisson regression

B) Quasi Maximum Likelihood estimation

"quasi-maximum likelihood" (quasi-ML) is a method used to estimate parameters in situations where the likelihood function is not strictly adhering to the assumptions of standard maximum likelihood estimation.

If you face **overdispersion** (variance is larger than predicted by a Poisson distribution), then quasi-ML can be used to account for this by using a different likelihood function or incorporating extra parameters into the model.