Introduction to Generalized Linear Models

Filippo Gambarota

University of Padova

2022/2023

Updated on 2023-04-25

Outline

1. Beyond the Gaussian distribution
2. Generalized Linear Models
3. Relevant distributions
4. Data simulation `#extra`

Beyond the Gaussian distribution

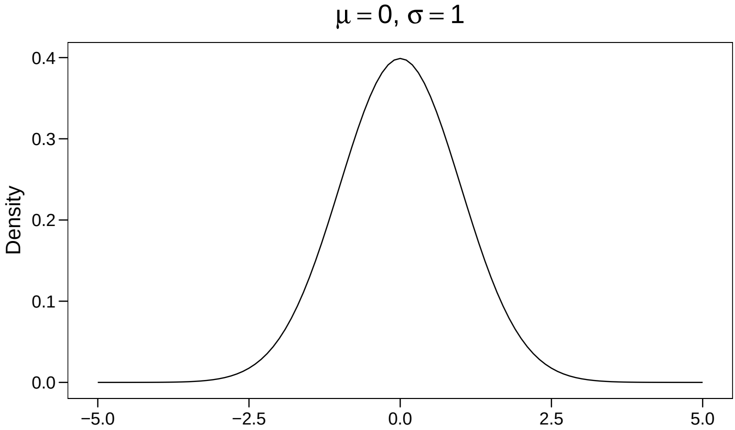
Quick recap about Gaussian distribution

- The Gaussian distribution is part of the Exponential family
- It is defined with mean (μ) and the standard deviation (σ) that are independent
- It is symmetric with the same value for mean, mode and median
- The support is $[-\infty, +\infty]$

The Probability Density Function (PDF) is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Quick recap about Gaussian distribution



But not always gaussian-like variables!

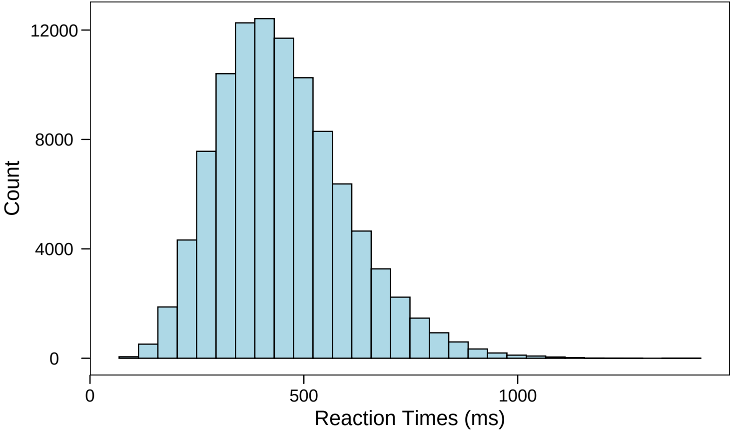
Quick recap about Gaussian distribution

In fact, in Psychology, variables do not always satisfy the properties of the Gaussian distribution. For example:

- Reaction times
- Accuracy
- Percentages or proportions
- Discrete counts
- Likert scales
- ...

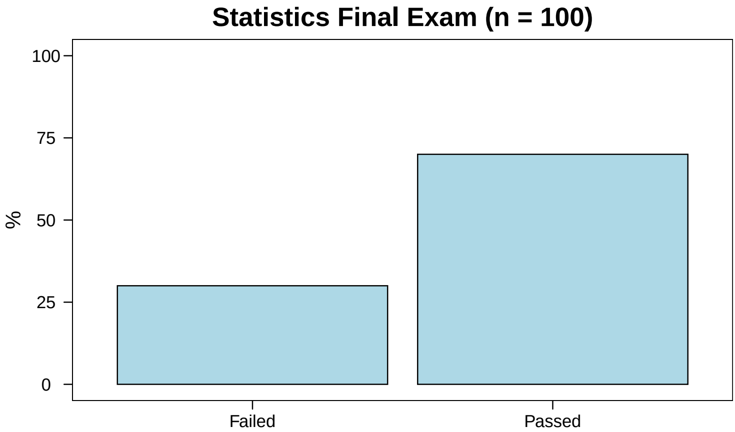
Reaction times

Measuring **reaction times during a cognitive task**. Non-negative and probably skewed data.



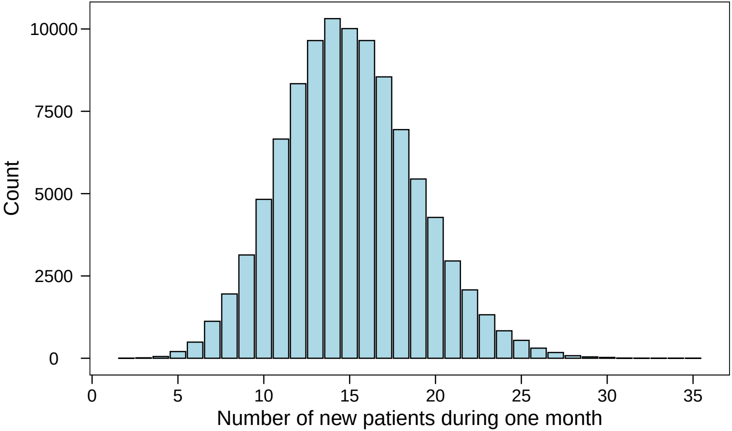
Binary outcomes

Counting the number of people passing the exam out of the total.
Discrete and non-negative. A series of binary (i.e., *bernoulli*) experiments.



Counts

Counting the number of new hospitalized patients during one month in different cities. Discrete and non-negative values.



Question...

Should we use a linear model for these variables?

Should we use a linear model for these variables?

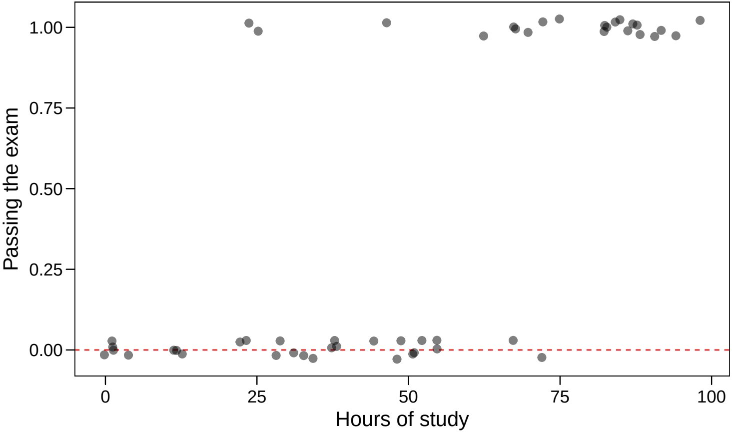
Let's try to fit a linear model on the probability of passing the exam ($N = 50$) as a function of the hours of study:

student	study.hours	passing
1	13	0
2	67	0
3	4	0
4	24	1
...	...	...
47	55	0
48	88	1
49	67	1
50	62	1

n	npassing	nfailing	ppassing
50	22	28	0.44

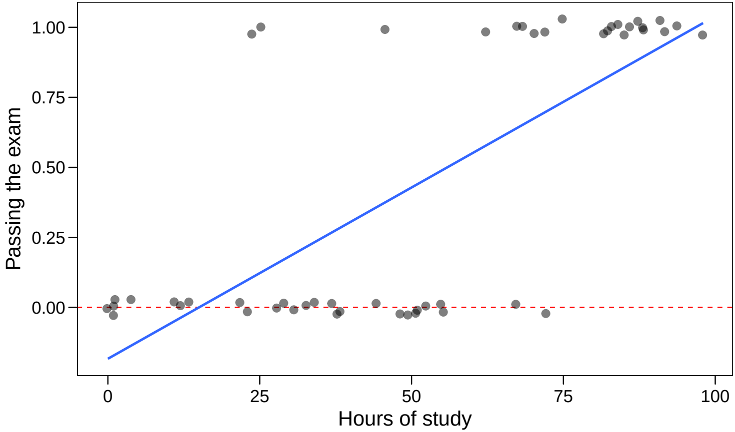
Should we use a linear model for these variables?

Let's plot the data:



Should we use a linear model for these variables?

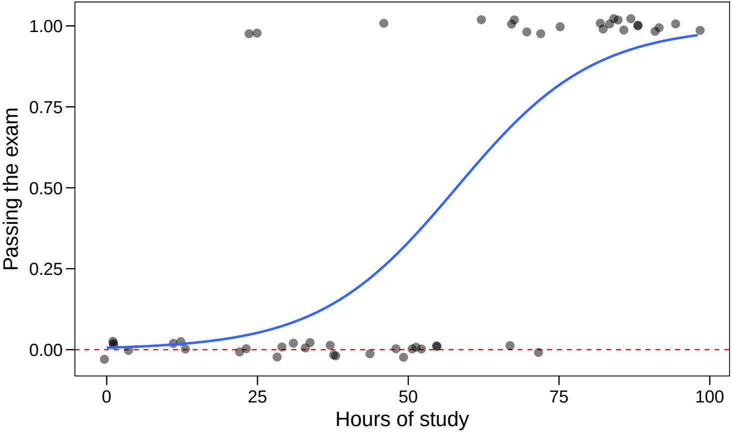
Let's fit a linear model `passing ~ study_hours` using `lm`:



Do you see something strange?

Should we use a linear model for these variables?

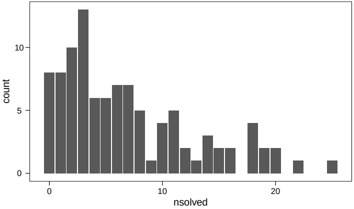
A little **spoiler**, the relationship should be probably like this:



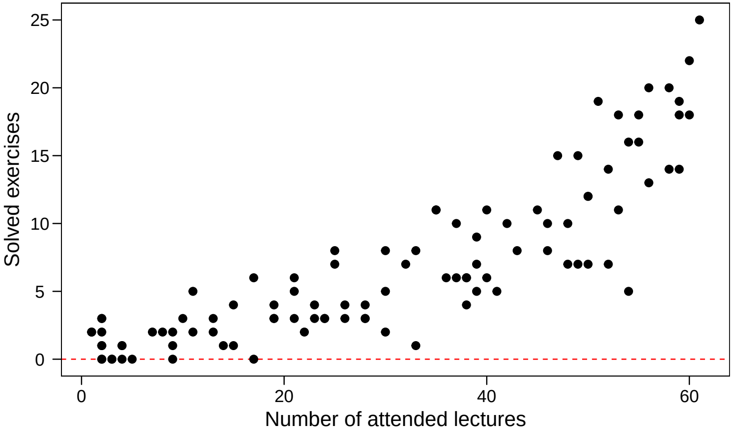
Should we use a linear model for these variables?

Another example, the number of solved exercises in a semester as a function of the number of attended lectures ($N = 100$):

student	attended.lectures	nsolved
1	15	4
2	35	11
3	56	20
4	61	25
...	...	...
97	9	1
98	37	10
99	54	16
100	3	0

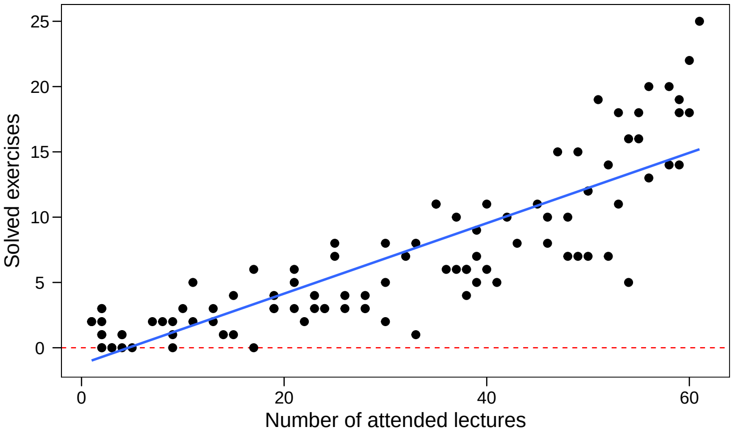


Should we use a linear model for these variables?



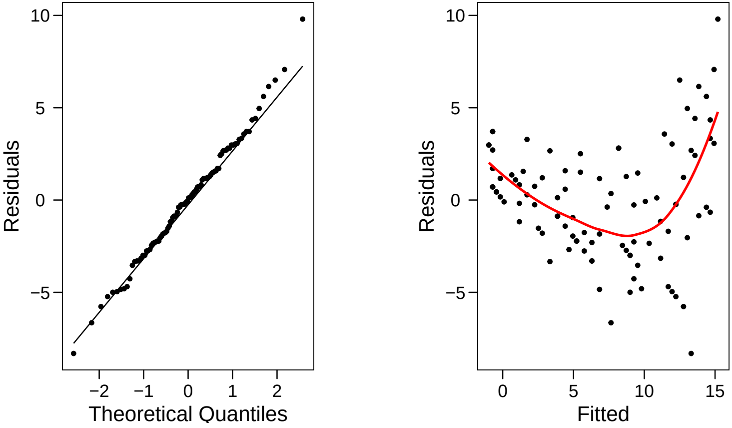
Should we use a linear model for these variables?

Again, fitting the linear model seems partially appropriate but there are some problems:



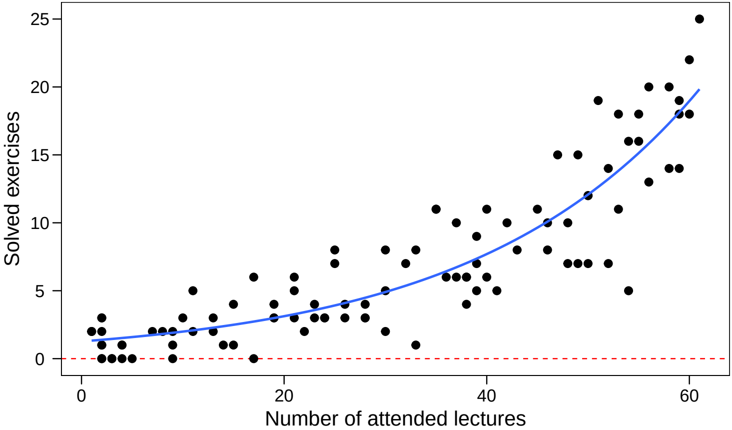
Should we use a linear model for these variables?

Also the residuals are quite problematic:



Should we use a linear model for these variables?

Another little spoiler, the model should consider both the support of the y variable and the non-linear pattern. Probably something like this:



So what?

Both linear models somehow capture the expected relationship but there are serious fitting problems:

- impossible predictions
- poor fitting for non-linear patterns

As a general rule in ~~life~~ statistics:

All models are wrong, some are useful.

— George Box

We need a new class of models...

- We need that our model take into account the **features of our response variable**
- We need a model that, **with appropriate transformation**, keep **properties of standard linear models**
- We need a model that is **closer to the true data generation process**

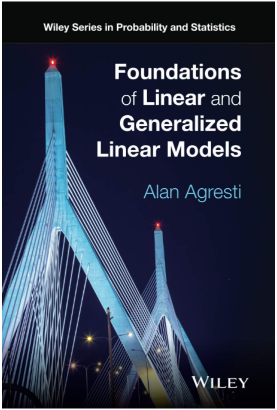
Let's switch to Generalized Linear Models!

Generalized Linear Models

Main references

For a detailed introduction about GLMs

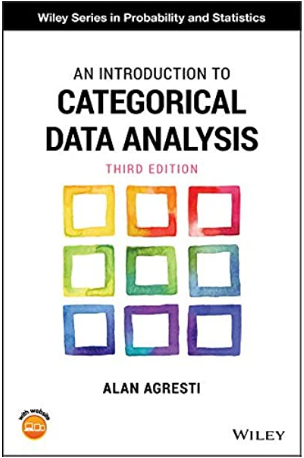
- Chapters: 1 (intro), 4 (GLM fitting), 5 (GLM for binary data)



Main references

For a basic and well written introduction about GLM, especially the Binomial GLM

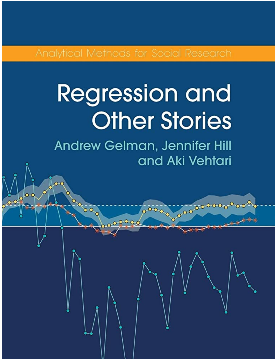
- Chapters: 3 (intro GLMs), 4-5 (Binomial Logistic Regression)



Main references

Great resource for interpreting Binomial GLM parameters:

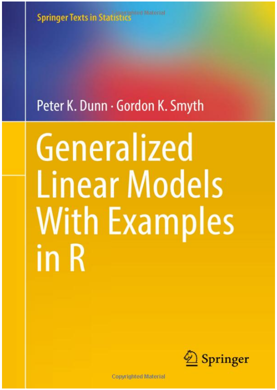
- Chapters: 13-14 (Binomial Logistic GLM), 15 (Poisson and others GLMs)



Main references

Detailed GLMs book. Very useful especially for the diagnostic part:

- Chapters: 8 (intro), 9 (Binomial GLM), 10 (Poisson GLM and overdispersion)



General idea

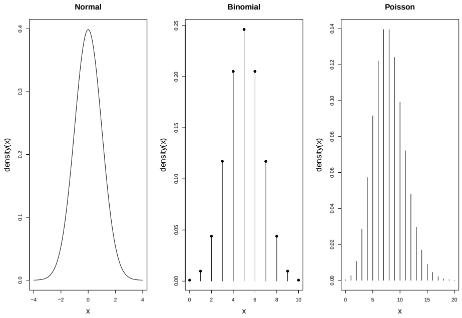
- models that assume distributions other than the normal distributions
- models that considers non-linear relationships

Recipe for a GLM

- **Random Component**
- **Systematic Component**
- **Link Function**

Random Component

The **random component** of a GLM identify the response variable Y and the appropriate probability distribution. For example for a numerical and continuous variable we could use a Normal distribution (i.e., a standard linear model). For a discrete variable representing counts of events we could use a Poisson distribution, etc.



Systematic Component

The **systematic component** or *linear predictor* (η) of a GLM is the combination of explanatory variables i.e. $\beta_0 + \beta_1 x_1 + ... + \beta_p x_p$.

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

When the **link function** (see next slide) is used, the relationship between η and the expected value μ of the **random component** is linear (as in standard linear models)

Link Function

The **link function** $g(\mu)$ is an **invertible** function that connects the expected value (i.e., the mean μ) of the probability distribution (i.e., the random component) with the *linear combination* of predictors $g(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$. The inverse of the link function g^{-1} map the linear predictor (η) into the original scale.

$$g(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$
$$\mu = g^{-1}(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)$$

Thus, the relationship between μ and η is linear only when the **link function** is applied i.e. $g(\mu) = \eta$.

Link function

The simplest **link function** is the **identity link** where $g(\mu) = \mu$ and correspond to the standard linear model. In fact, the linear regression is just a GLM with a **Gaussian random component** and the **identity** link function.

There are multiple **random components** and **link functions** for example with a 0/1 binary variable the usual choice is using a **Binomial** random component and the **logit** link function.

Family	Link	Range
gaussian	identity	$(-\infty, +\infty)$
binomial	logit	$\frac{0,1,\ldots,n_i}{n_i}$
	probit	$\frac{0,1,\ldots,n_i}{n_i}$
poisson	log	$0, 1, 2, \ldots$

Relevant distributions

Binomial distribution

The probability of having k success (e.g., 0, 1, 2, etc.) out of n trials with a probability of success p is:

$$f(n, k, p) = Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

The np is the mean of the binomial distribution and $np(1 - p)$ is the variance.

Bernoulli distribution

The **binomial** distribution is just a repetition of k **Bernoulli** trials. A single Bernoulli trial is:

$$
\begin{aligned}
f(x, p) &= p^x (1-p)^{1-x} \\
x &\in \{0, 1\}
\end{aligned}
$$

The mean is p and the variance is $p(1-p)$

Bernoulli and Binomial

The simplest situation for a Bernoulli trial is a coin flip. In R:

```
n <- 1
p <- 0.7
rbinom(1, n, p) # a single bernoulli trial
```

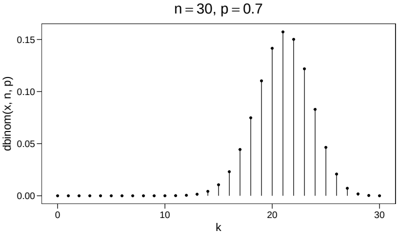
```
## [1] 1
```

```
n <- 10
rbinom(10, 1, p) # n bernoulli trials
```

```
##  [1] 1 1 1 1 1 1 0 1 1 0
```

```
rbinom(1, n, p) # binomial version
```

```
## [1] 4
```



Bernoulli and Binomial

The Bernoulli and the Binomial distributions are used as **random components** when we have the dependent variable assuming 2 values (e.g., *correct* and *incorrect*) and we have the total number of trials:

- Accuracy on a cognitive task
- Patients recovered or not after a treatment
- People passing or not an exam

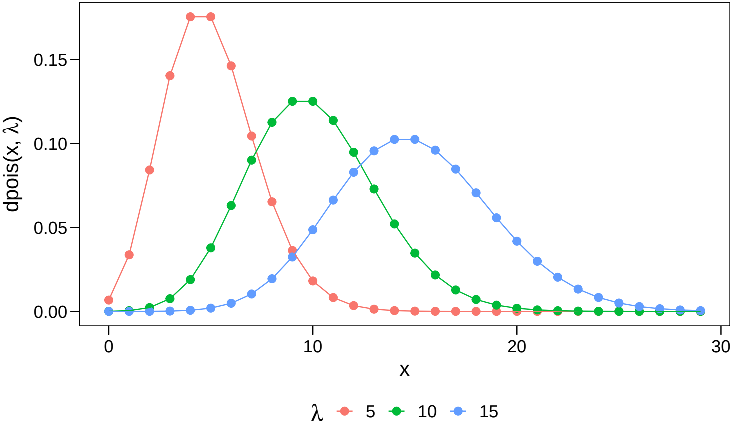
Poisson distribution

The number of events k during a fixed time interval (e.g., number of new user on a website in 1 week) is:

$$f(k, \lambda) = Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where k is the number of occurrences ($k = 0, 1, 2, ...$), e is Euler's number ($e = 2.71828...$) and ! is the factorial function. The mean and the variance of the Poisson distribution is λ

Poisson distribution



As λ increases, the distribution is well approximated by a Gaussian distribution, but the Poisson is discrete.

Data simulation #extra

Data simulation #extra

- During the course we will try to simulate some data. Simulating data is an amazing education tool to understand a statistical model.
- By simulating from a **generative model** we are doing a so-called **Monte Carlo Simulations** [1]

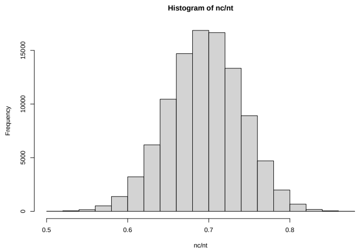
Data simulation `#extra`

In R there are multiple functions to generate data from probability distributions:

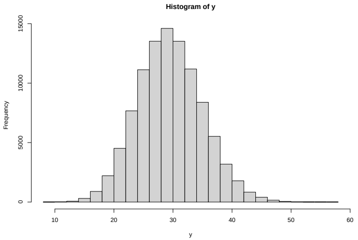
Function	Distribution	Action
`d`	norm	Compute the density
	pois	
	binom	
`p`	norm	Return the cumulative probability given a quantile
	pois	
	binom	
`q`	norm	Return the quantile given a cumulative proability
	pois	
	binom	
`r`	norm	Generate random numbers
	pois	
	binom	

Data simulation #extra

```r
n <- 1e5 # number of experiments
nt <- 100 # number of subjects
p <- 0.7 # probability of success
nc <- rbinom(n, nt, p)
```




```r
n <- 1e5 # number of subjects
lambda <- 30 # mean/variance
y <- rpois(n, lambda)
```



References

[1] J. E. Gentle, "Monte carlo methods for statistical inference," in *Computational statistics*, J. E. Gentle, Ed., New York, NY: Springer New York, 2009, pp. 417–433. doi: 10.1007/978-0-387-98144-4_11.

filippo.gambarota@unipd.it

github.com/filippogambarota