Are all interaction false?

The importance of the appropriate distribution and link function with non-normal data

Filippo Gambarota[1] Enrico Toffalini[2]

[1]Department of Developmental Psychology and Socialization
University of Padova

[2]Department of General Psychology
University of Padova

@AIP Psicologia dello Sviluppo e dell'Educazione 2024

When the effect of one variable x is supposed to be **moderated** from another variable z and your **response variable y is non-normal**, the interaction is probably a **false positive**.

Beyond the Gaussian distribution

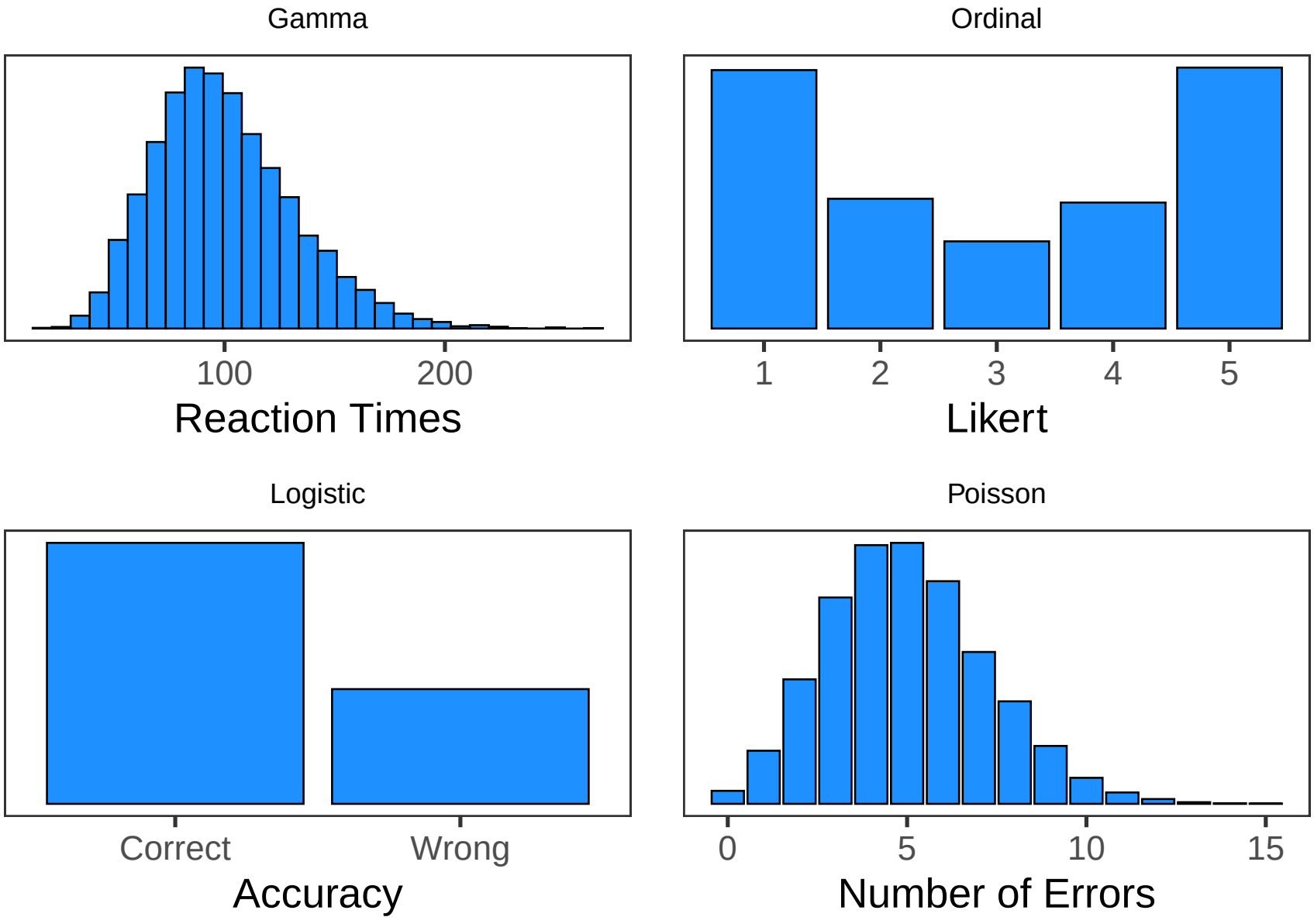
Beyond the Gaussian distribution

Warning

In Psychology, most of the time we deal with non-normal distributions due to the measure that we are using or the type of variable.

- **Time**: response times, reading times, etc.
- **Counts**: number of errors, number of symptoms, sum of likert items
- ...

Beyond the Gaussian distribution



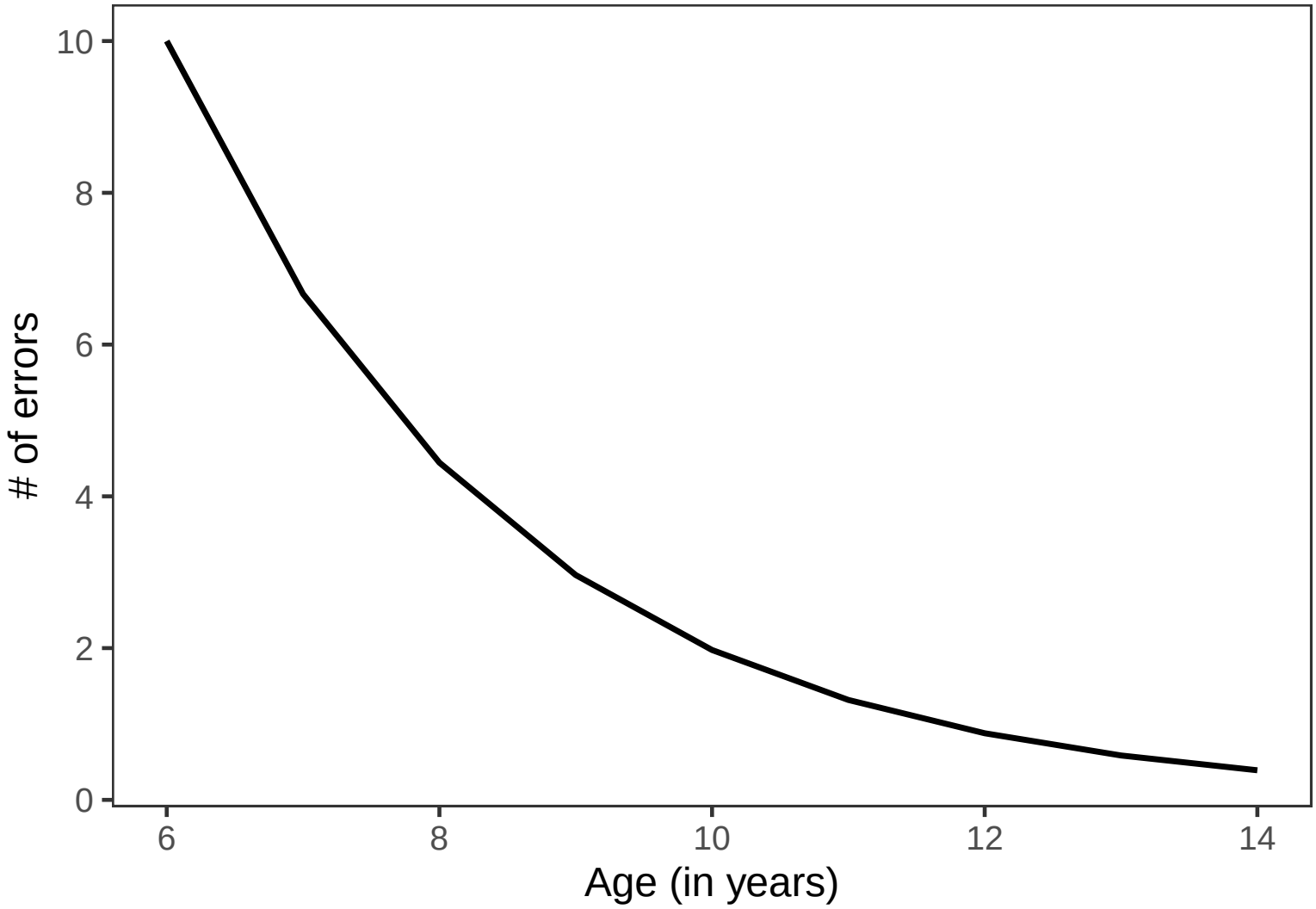
Are interactions important?

Beyond main effects...

Most of the time in Psychology, **we are not really interested in the main effects**, but how the effect of a focal variable x (e.g., treatment) on the response variable y is **moderated by another variable** z (e.g., age).

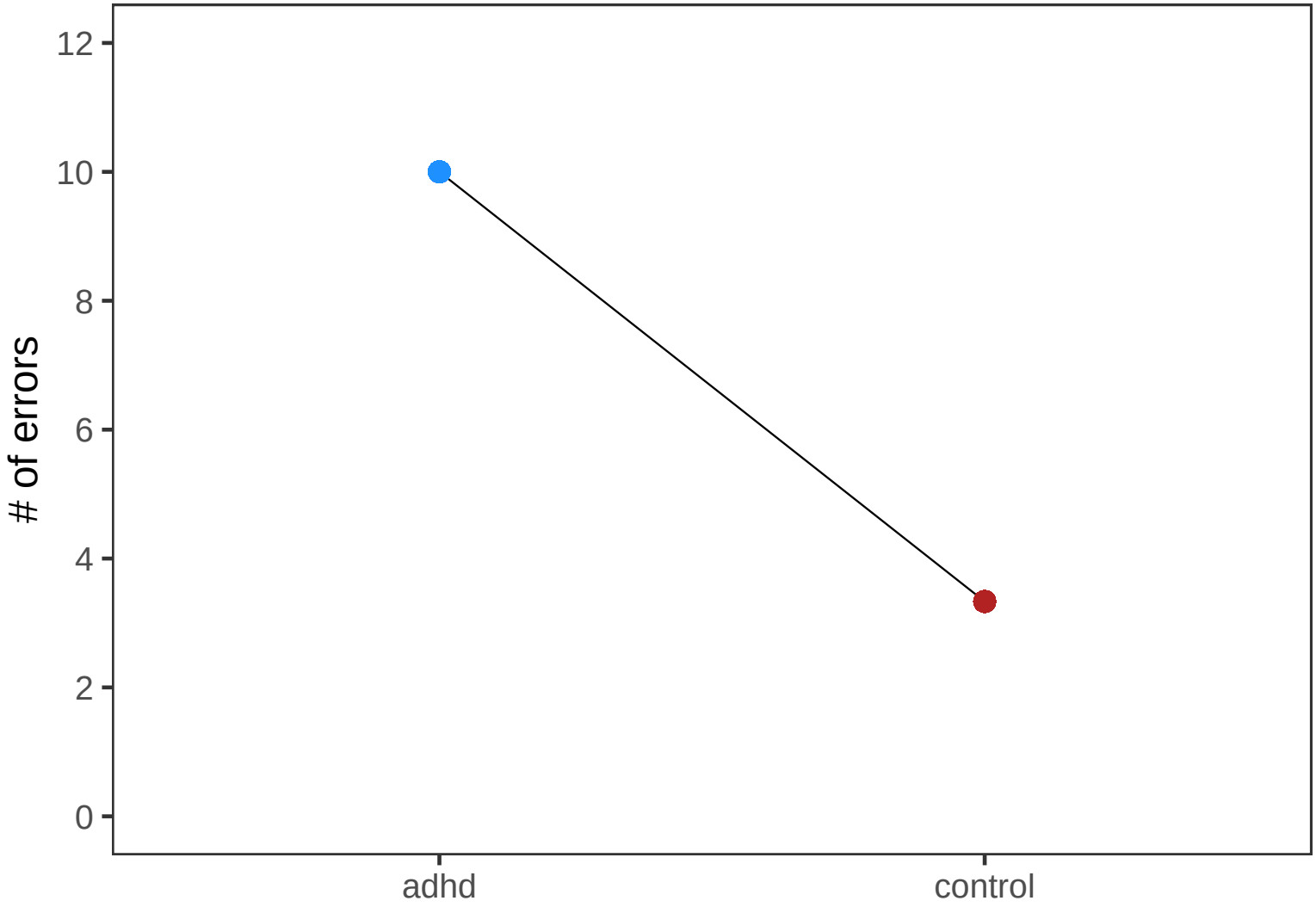
An example...

We are evaluating the effect of **age** on the number of **errors** during a task. We expect that older children commit a lower number of errors.



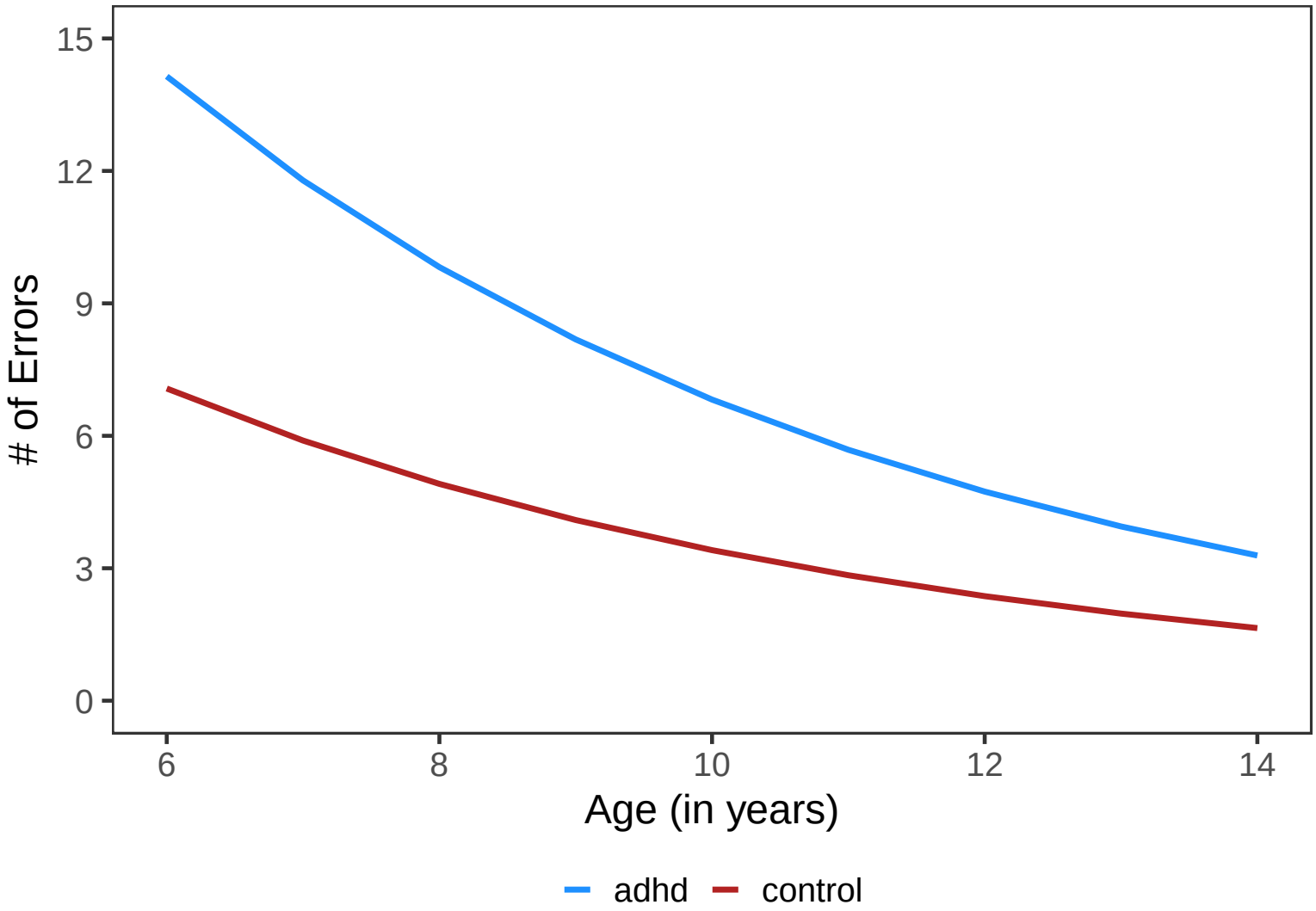
An example...

Similarly, we could compare a clinical group (e.g., ADHD) and a control group expecting more errors in the former.



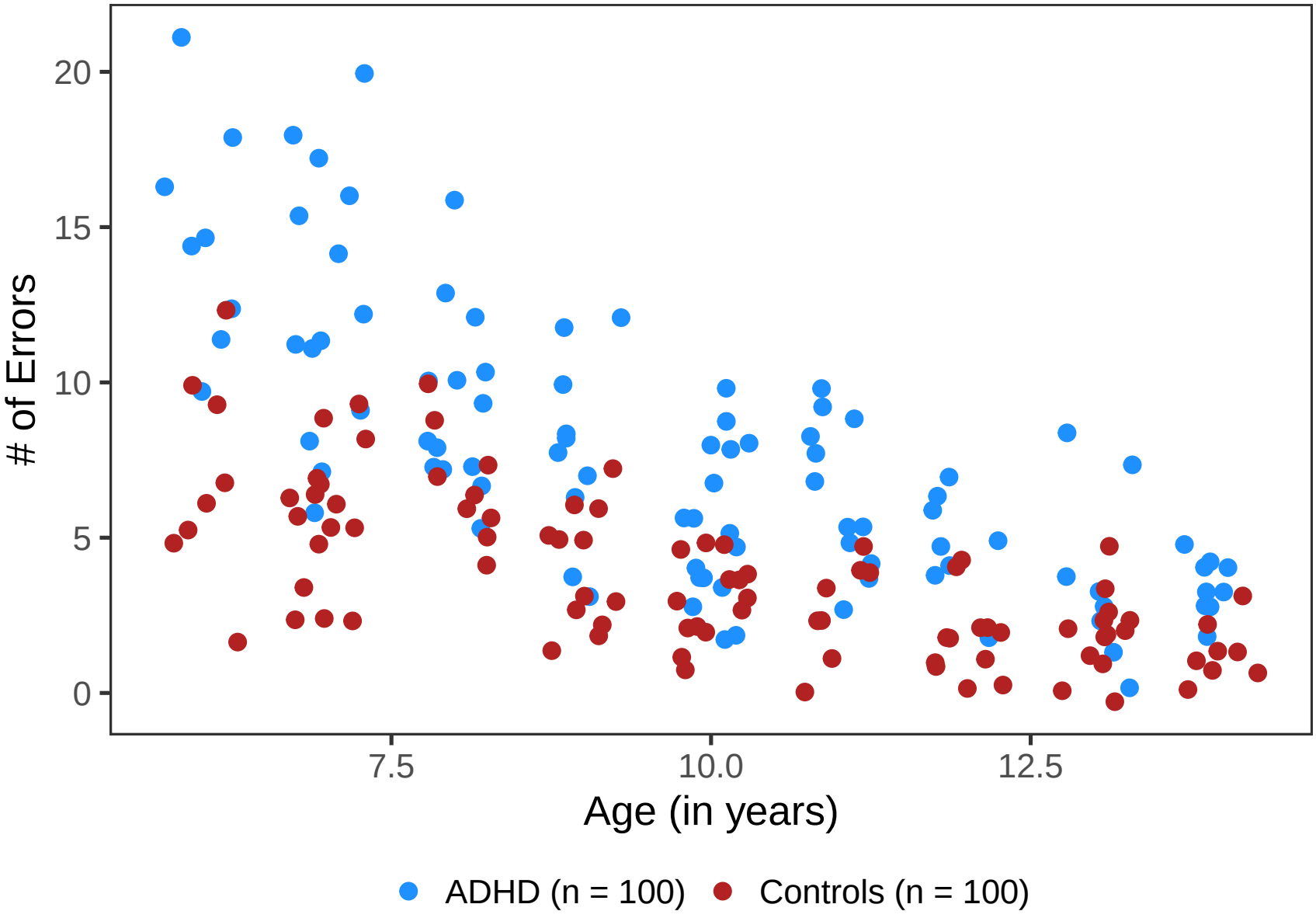
An example...

Usually, what we are really interested is the interaction. Thus how the age effect change according to the group.

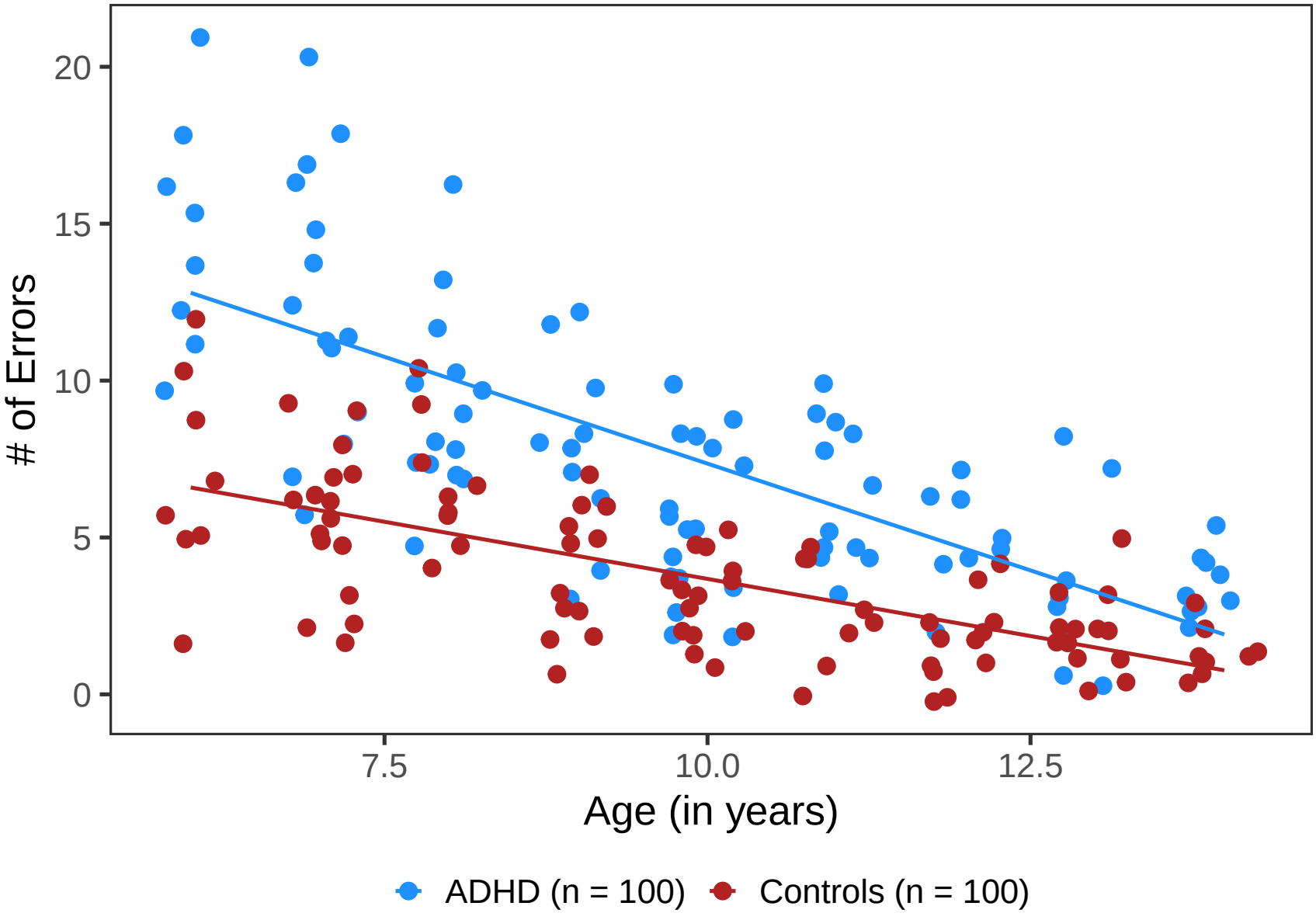


A little quiz!

An example with real data...



Is there (graphical) evidence for interaction?



The linear model results

In the previous plot we fitted a standard linear model predicting the number of errors with **age**, **group** and the **interaction**.

Effect	**SS**	**F**	**p**
Age	1273.651	199.020	< 0.001
Group (adhd vs controls)	711.473	111.175	< 0.001
Age x Group	119.496	18.672	< 0.001

The model suggest evidence for an interaction effect, PAPER ACCEPTED!

The linear model has been scammed!

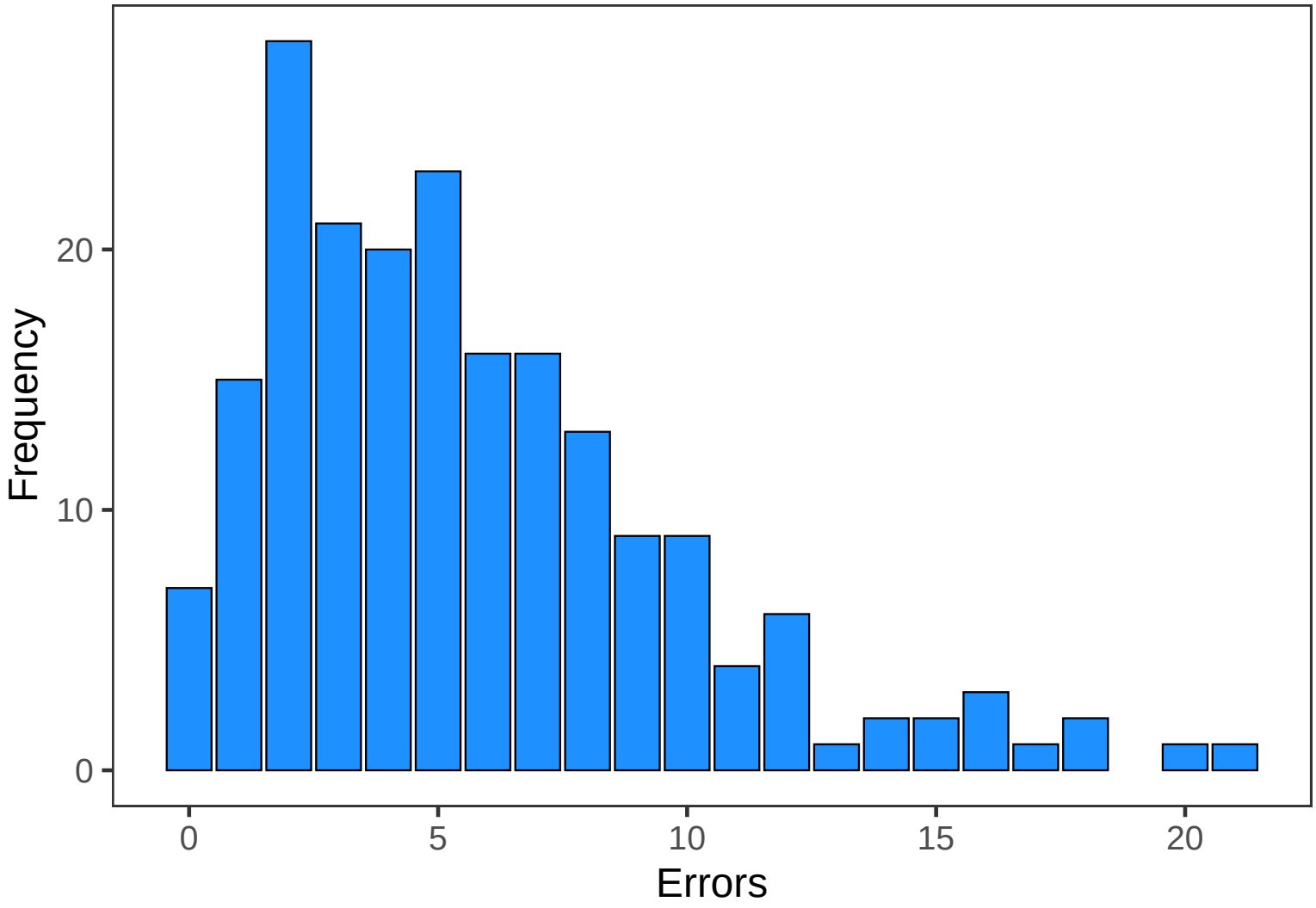
In reality, the previous dataset has been simulated. And this is (roughly) the generative model:

$$y_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{group}_i + \beta_3 \text{age}_i \text{group}_i$$

But the β_3 parameter (i.e., the interaction) has been fixed to 0. In other words, **there is no interaction**.

Why?

The main reason is that **errors** is a discrete variable bounded between 0 and $+\infty$.

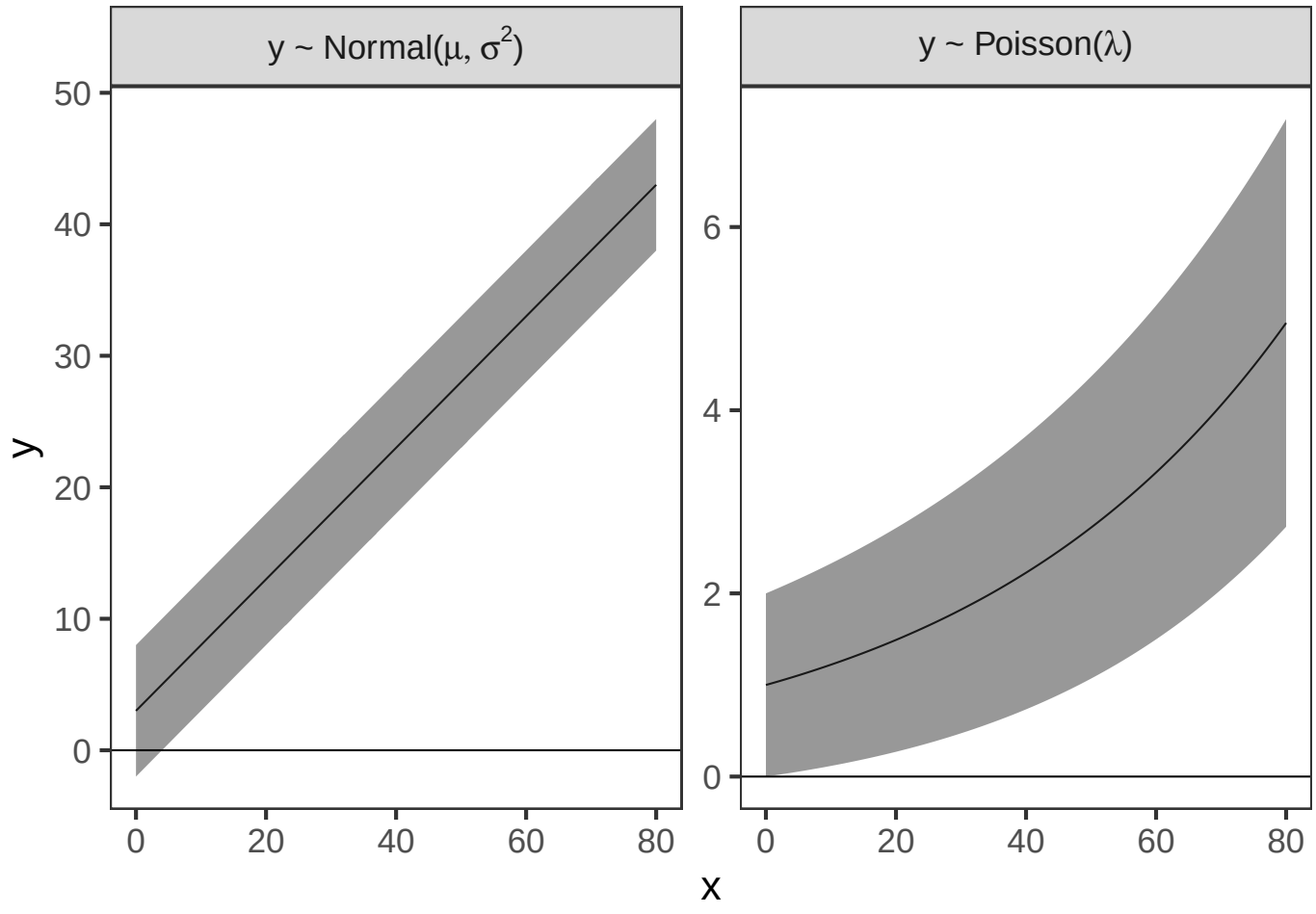


Why this is a problem?

The linear model (t-test, regression, etc.) is not aware of this relationship. The model fit straight lines ignoring the type of variable, the presence of bounds and the mean-variance relationship.

Beyond the normal distribution, Poisson!

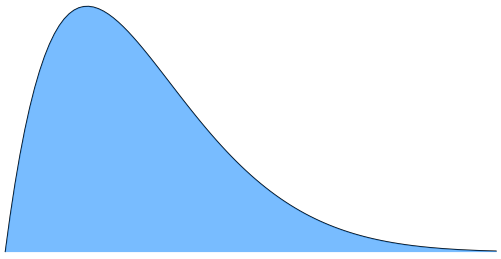
Mean and variance are linked in the Poisson distribution because there is a lower bound.



We need a “new” class of models!

Generalized linear models, the big picture

Non-normal data

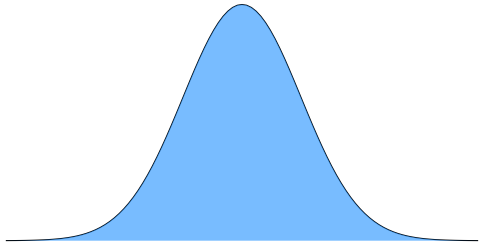


Distribution (e.g., Poisson)

Link Function (e.g, log)

$$\eta = g(y)$$

Link function space



Fitting a standard linear model

$$\boldsymbol{\eta} \sim \boldsymbol{\beta X}$$

- estimating $\boldsymbol{\beta}_j$
- hypothesis testing
- ...

Inverse Link Function (e.g, exp)

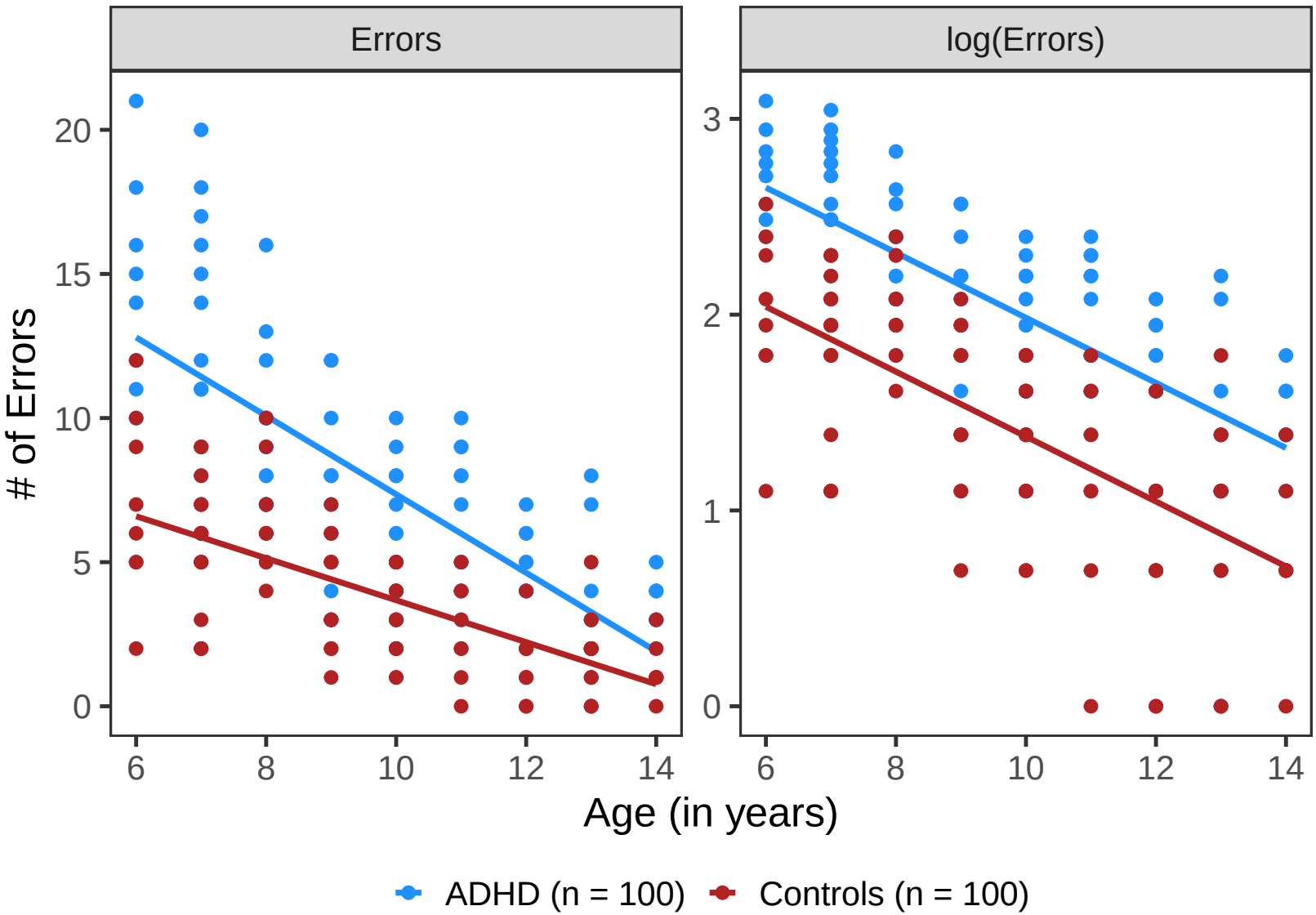
$$y = g^{-1}(\eta)$$

Reporting

- Intepreting results
- plotting
- effect size
- ...

Poisson regression and log link function

For the Poisson, the usual link function is the **logarithm**, that *stabilize the mean-variance relationship*.



Not so "new", but rarely used in psychology

J. R. Statist. Soc. A,
(1972), **135**, *Part* 3, p. 370

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

Rothamsted Experimental Station, Harpenden, Herts

SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

Not so “new”, but rarely used in psychology

J. R. Statist. Soc. A,
(1972), **135**, *Part* 3, p. 370

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

Rothamsted Experimental Station, Harpenden, Herts

SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

GLM are easy in R (and in other software)

In R (but also in other software) we can just switch from the `lm` to the `glm` function. We only need to specify the **distribution** and the **link function** to use.

```
fit_lm <- lm(errors ~ group * age0)
fit_glm <- glm(errors ~ group * age0,
               family = poisson(link = "log"))
```

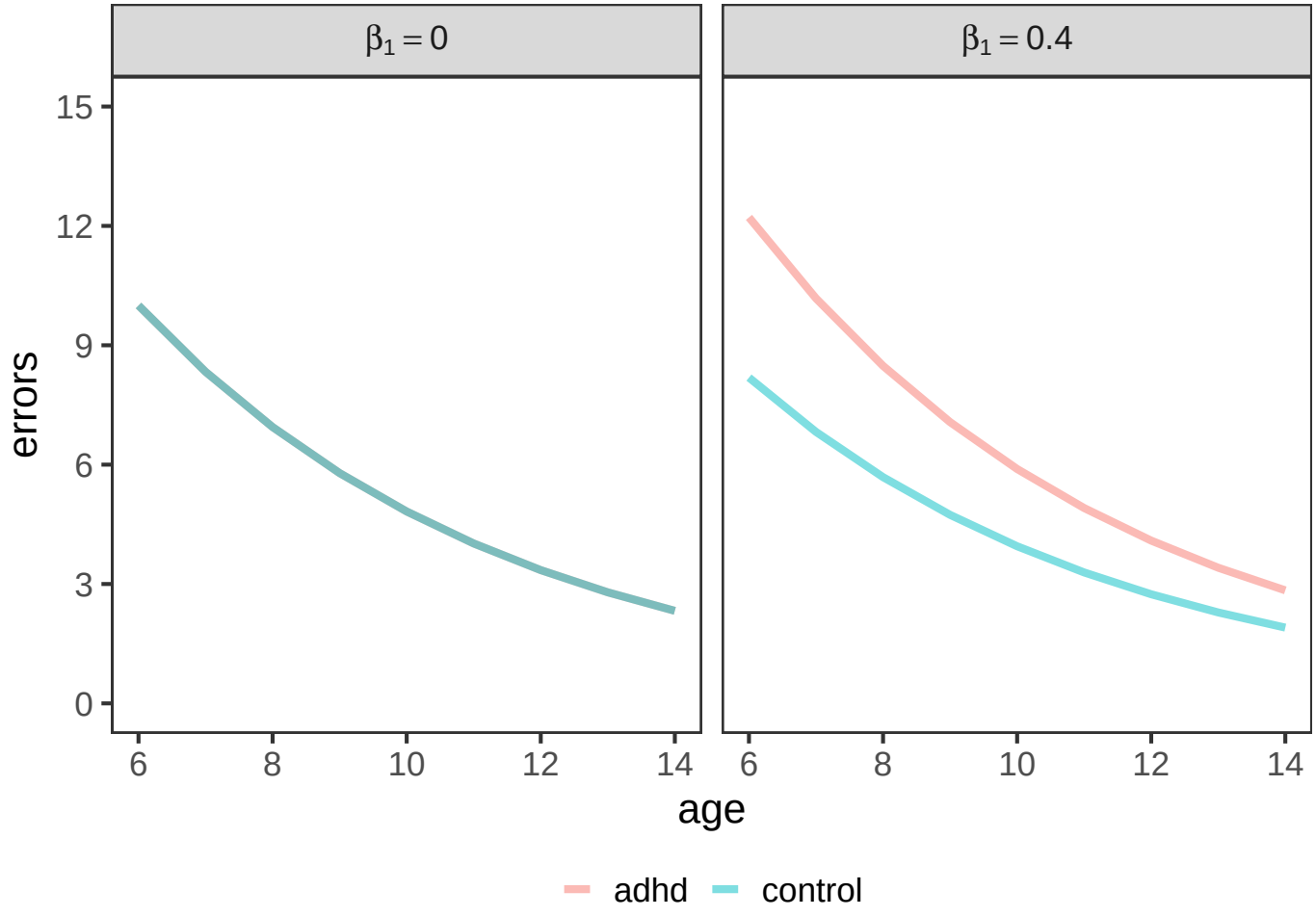
GLM results

When using the GLM, the interaction is no longer significant. **The linear model was commiting type-1 error**.

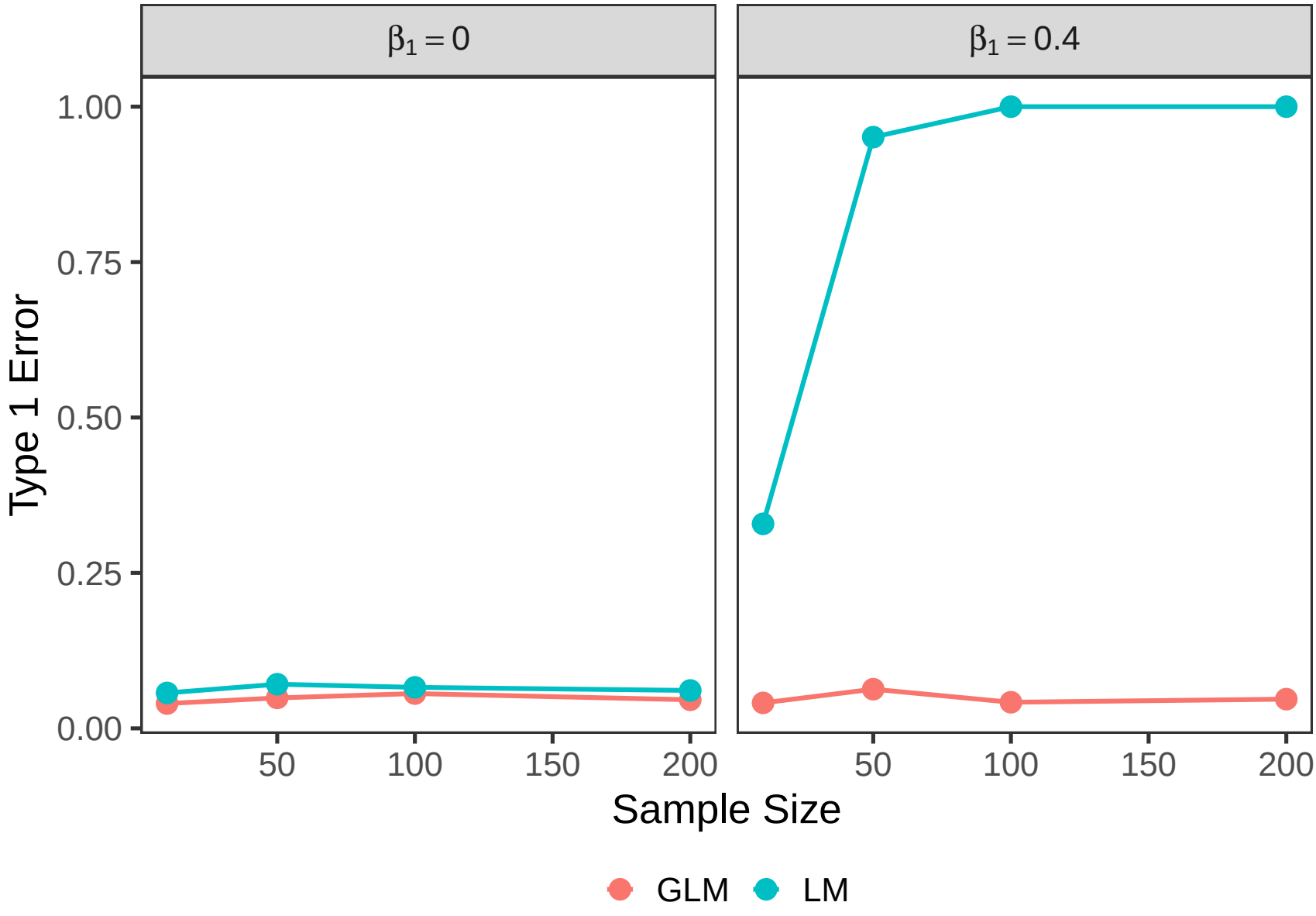
Effect	χ^2	**p**
Age	130.685	< 0.001
Group (adhd vs controls)	235.890	< 0.001
Age x Group	0.341	0.559

How serious is the problem?

This is the simulation setup. In both cases the interaction is fixed to 0 and we simulated different sample sizes $n = [10, 50, 100, 200]$ and we used the linear model and the GLM.



Very serious!



filippo.gambarota@unipd.it

filippogambarota.github.io