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

Are interactions important?

Main effects...

Overall, is the treatment effective?

Overall, is there a difference between the clinical and the control group?

Interactions...

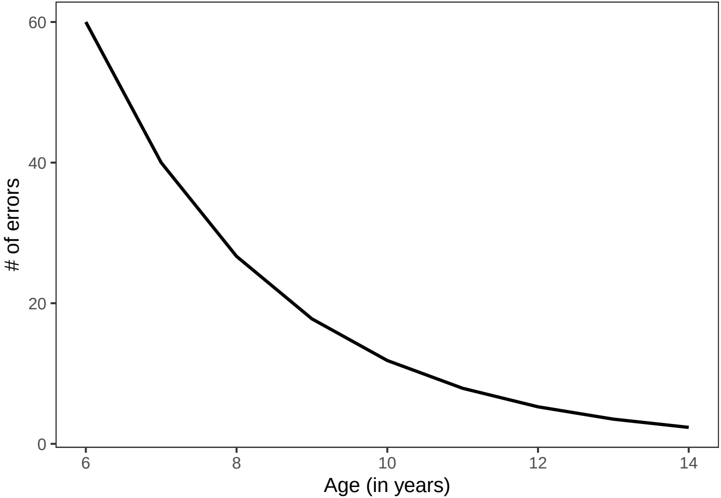
Often, we are more interested in testing interaction effects:

Is the treatment effective? in the condition A vs B?

Is the difference between the clinical and control group higher for older/younger children?

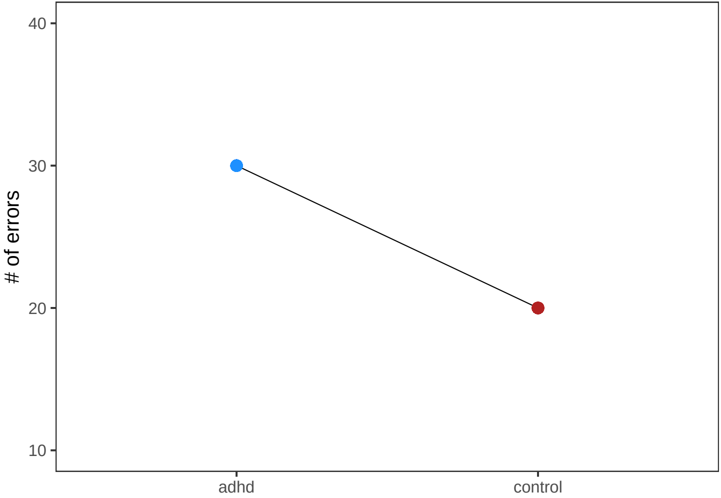
Are interactions important?

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.



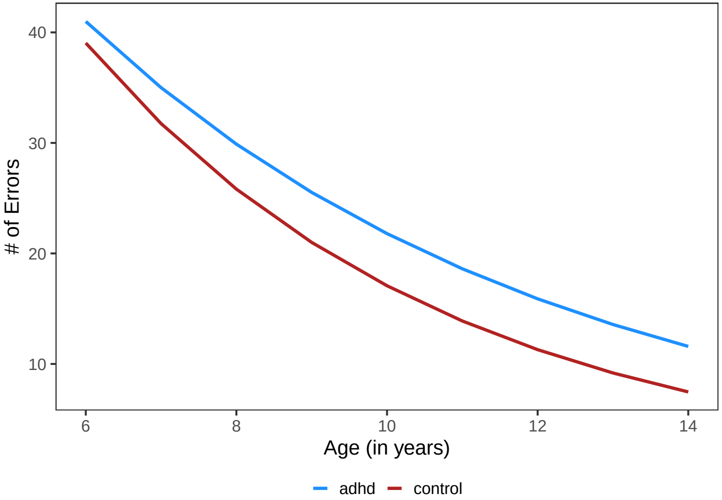
Are interactions important?

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



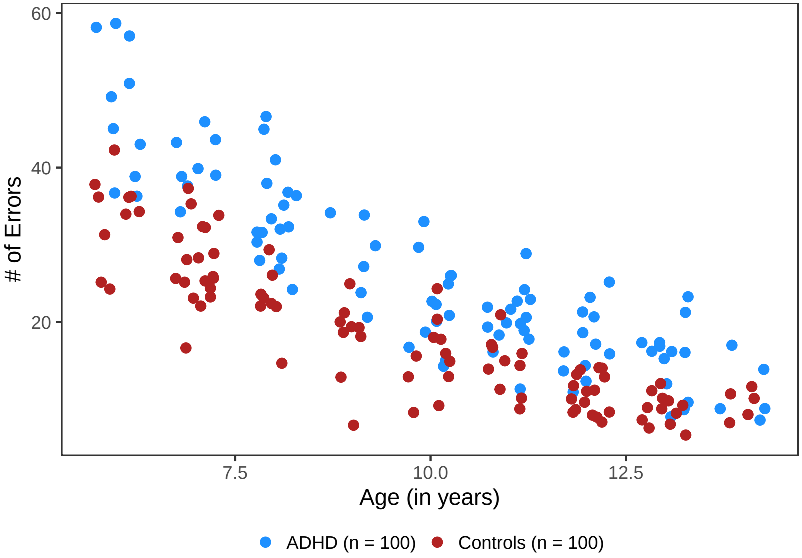
Are interactions important?

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

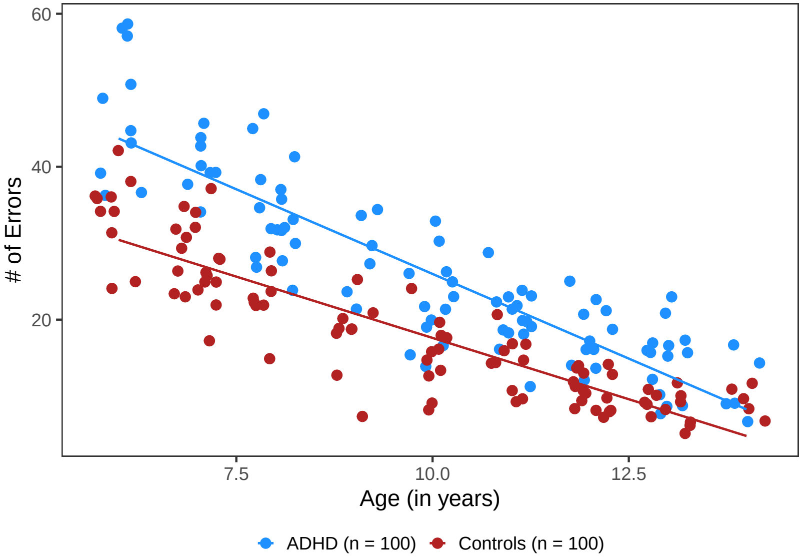


A little quiz!

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**.

```
lm(formula = errors ~ age0 * group, data = dd)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    43.6889     0.9878  44.228  < 2e-16 ***
age0                           -4.4279     0.2130 -20.790  < 2e-16 ***
groupControls (n = 100)       -13.2559     1.3503  -9.817  < 2e-16 ***
age0:groupControls (n = 100)    1.2224     0.2947   4.147 5.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The linear model has been scammed!

In reality, the previous dataset has been simulated. And this is (roughly)[1] the generative model:

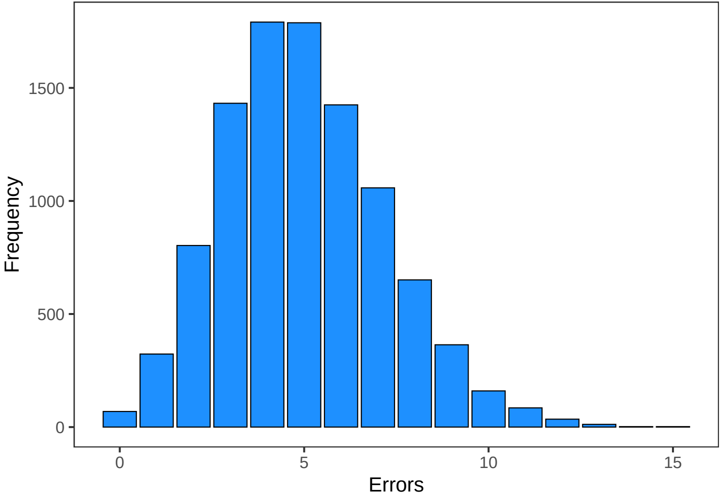
$$y_i = \beta_0 + \beta_1 \mathsf{age}_i + \beta_2 \mathsf{group}_i + \beta_3 \mathsf{age}_i \mathsf{group}_i$$

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

[1] *Roughly* because data are not generated by a standard linear model, see the next slides.

Why?

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



Beyond the normal distribution, Poisson!

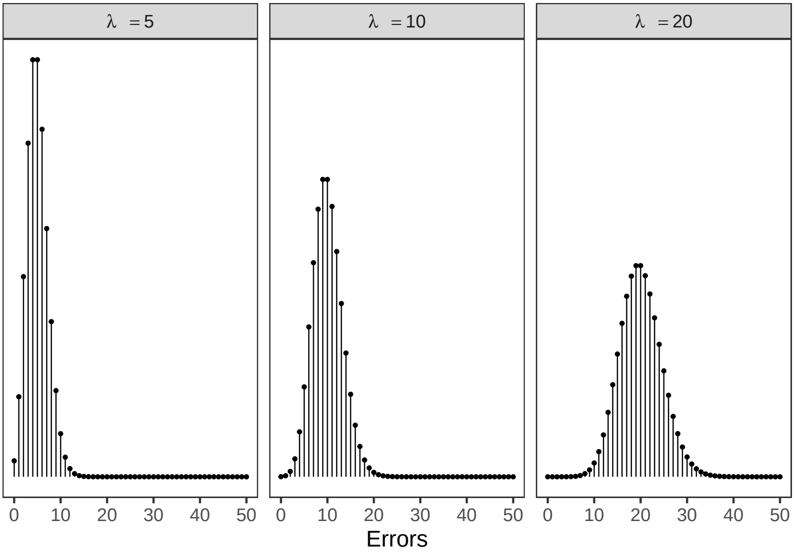
Beyond the specific equation, **mean and variance are linked** (in fact are the same value). This is completely different from the Gaussian distribution.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$E(X) = Var(X) = \lambda$$

Beyond the normal distribution, Poisson!

As the mean increase, also the variance increase!

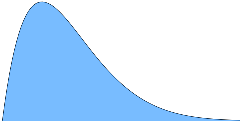


Why this is a problem?

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.

Generalized linear models, the big picture

Non-normal data

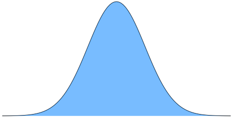


Distribution (e.g., Poisson)

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)

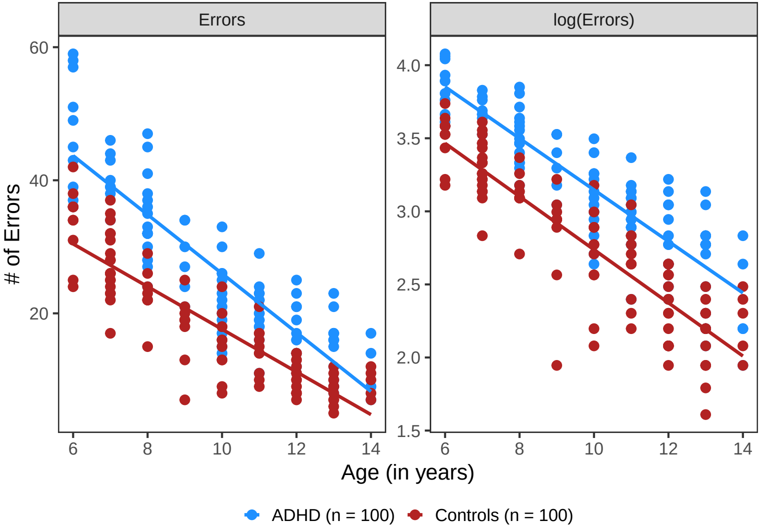
$$y = g^{-1}(\eta)$$

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*.



GLM are easy in R

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.

```
glm(formula = errors ~ group * age0, family = poisson(link = "log"),
    data = dd)
groupControls (n = 100)      -0.363567   0.047618  -7.635 2.26e-14 ***
age0                         -0.171730   0.008403 -20.436  < 2e-16 ***
groupControls (n = 100):age0 -0.011031   0.013047  -0.845    0.398
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

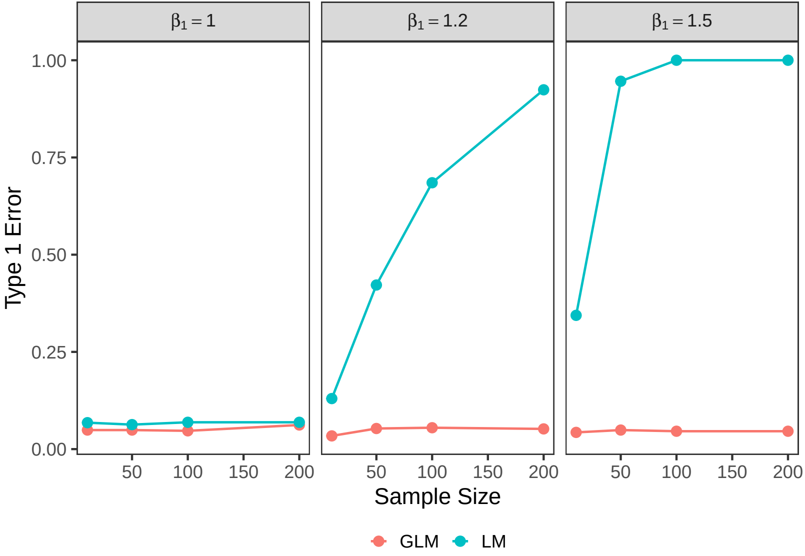
How serious is the problem?

We simulated the same scenario with different main effects of the **group** (i.e., the difference on average between ADHD and controls) and sample sizes.

- $\beta_1 = [0, 0.18, 0.4]$
- $n = [10, 50, 100, 200]$

We fitted the LM and the GLM and checked if the p value of the interaction is lower than α. We repeated each simulation 1000 times and calculated the proportion of false H_0 rejections.

Very serious!



Take-home message

- With non-normal data, using a standard linear model increase the type-1 error rate for the interaction term

Take-home message

- With non-normal data, using a standard linear model increase the type-1 error rate for the interaction term
- The type-1 error inflation increase as the main effect and the sample size increase

Take-home message

- With non-normal data, using a standard linear model increase the type-1 error rate for the interaction term
- The type-1 error inflation increase as the main effect and the sample size increase
- In some conditions, the type-1 error rate is above 50%. Thus more than 50% of our conclusions are false positives

Take-home message

- With non-normal data, using a standard linear model increase the type-1 error rate for the interaction term
- The type-1 error inflation increase as the main effect and the sample size increase
- In some conditions, the type-1 error rate is above 50%. Thus more than 50% of our conclusions are false positives
- Using a GLM with the appropriate distribution and link function controls the type-1 error rate in all simulated conditions

filippo.gambarota@unipd.it

filippogambarota.github.io