# Generalized Linear Model

**Linear Models and Mixed Models for categorical and count dependent variables**

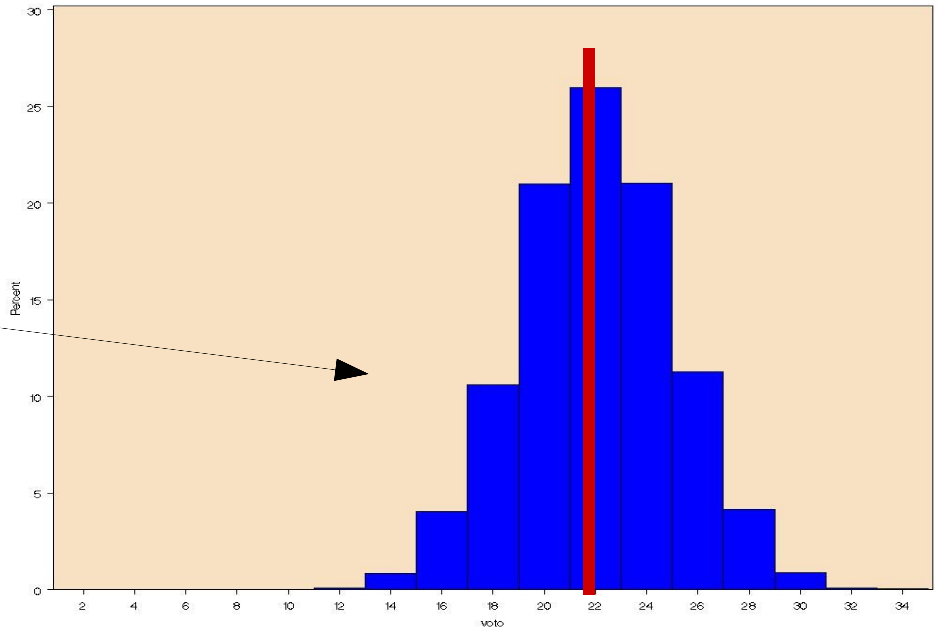
Marcello Gallucci
Univeirsity of Milano-Bicocca

# GLM Assumptions

$$y_i = a + e_i$$

$$corr(e_i, e_j) = 0$$

**3) Random variations are normally distributed**




$$e_i \sim N(0, \sigma)$$

# Generalized Linear Models

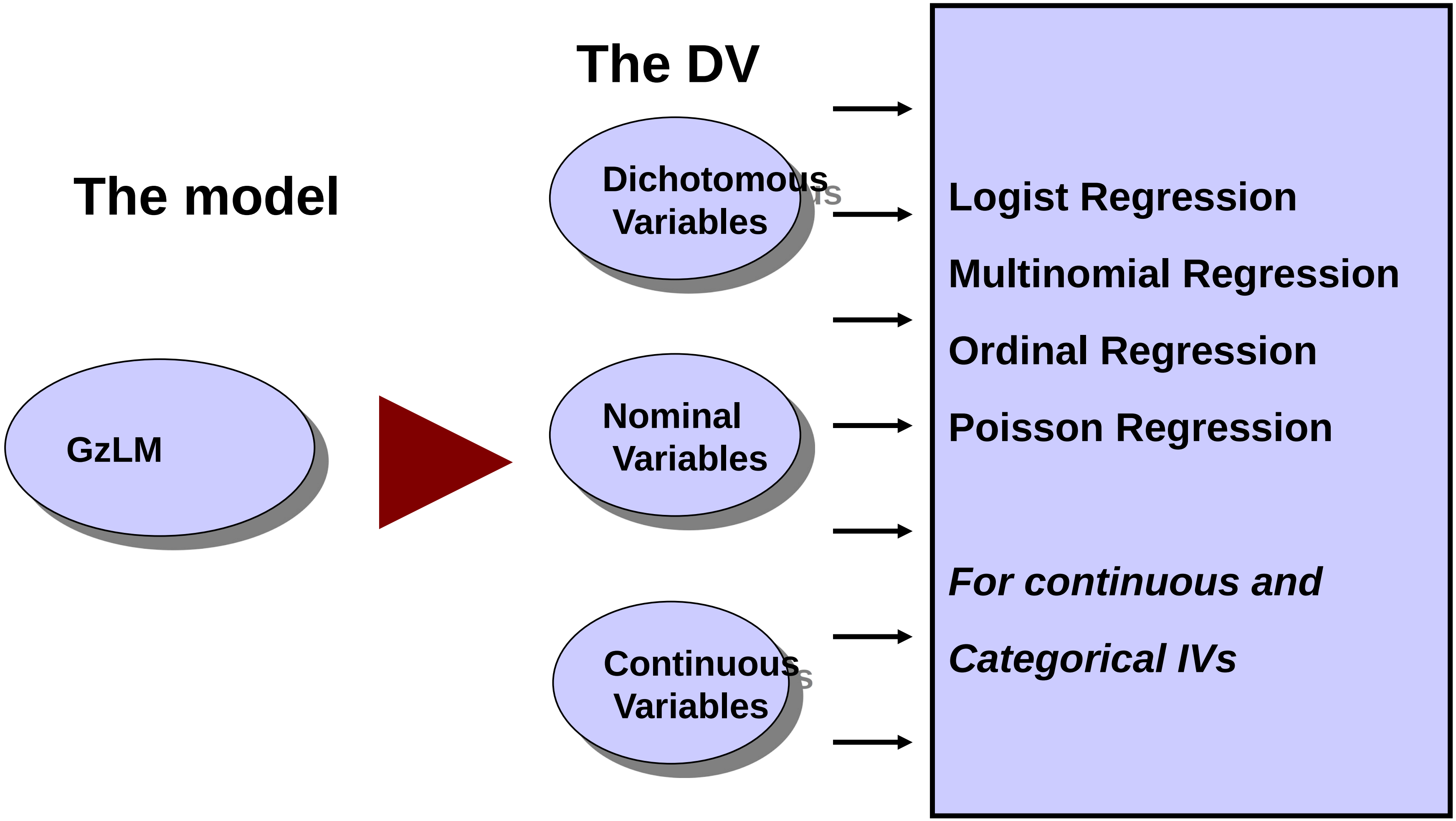
- There are many situations where the dependent variable is not normally distributed:
    - Predicting groups
    - Predicting  choices (yes/no, left/right, etc.)
    - Predicting frequencies of behavior

# GLM

When the assumptions are NOT met because the dependent variable is not normally distributed (dichotomous, frequencies, categorical etc), we generalize the GLM to the

Generalized Linear Model (GzLM)

# Generalized Linear Model



# GGLM

- The generalized linear model is a linear model with the dependent variable modelled with a specific function (link function) and with specific error distribution

**Generalized Linear Model**

$$f(y_i) = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + .. b_k \cdot x_{ki} + e_i$$

**Dependent variable**

**Specify a distribution shape**

# Generalized linear model

- Applying this logic we obtain a large set of possible statistical techniques

$$f(y_i) = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + .. b_k \cdot x_{ki} + e_i$$

| Dependent Variable | function | Distribution |
|---|---|---|
| Continuous | identity | Normal |
| Dichotomous | Logit of odd | Binomial |
| Categorical | Logit of odd | Multinomial |
| Ordinal | Cumulative Logit | Multinomial |
| Frequencies | Frequencies LN | Poisson |

# Generalized Linear Model:

# Logistic model

# The Logistic Regression

- The aim of Logistic regression is to estimate the effects of one or more IV on a dichotomous dependent variable

- Logistic regression is a particular case of the Generalized(General)LM

- Due to our knowledge of the GLM, we can apply all the techniques of the GLM (regression, ANOVA, interactions, etc.) to the case of dichotomous dependent variable

- To understand logistic regression, it is useful to understand why we cannot use the GLM (linear regression) as we already know it

# Linear regression assumptions

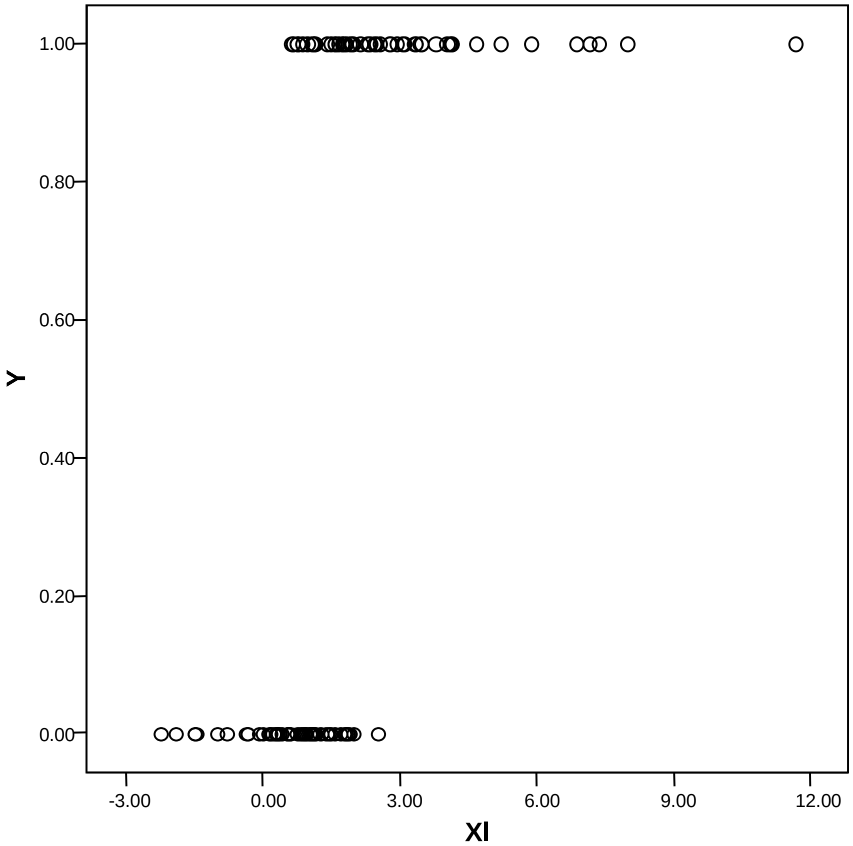
- When we run a GLM model (regression, ANOVA, etc) we are implicitly making specific assumptions:

| What we do | Corresponding assumption |
| --- | --- |
| Estimates the effects | The effect is linear or conditional linear |
| Estimates variance explained | The error variance is constant across the predicted values |
| Test for significance | The errors of the predicted values are normally distributed |

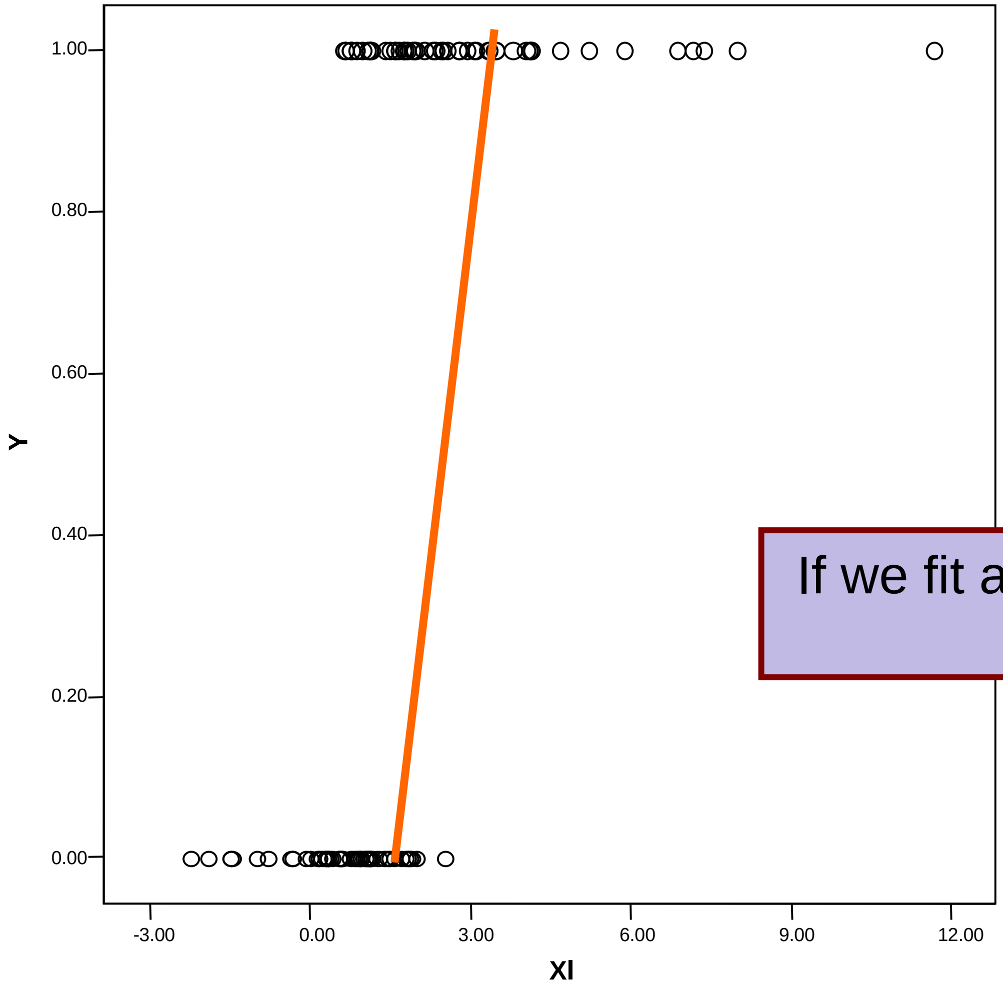
...

# Example

- When the DV is dichotomous, the scatterplot Y~X will always look something like this:



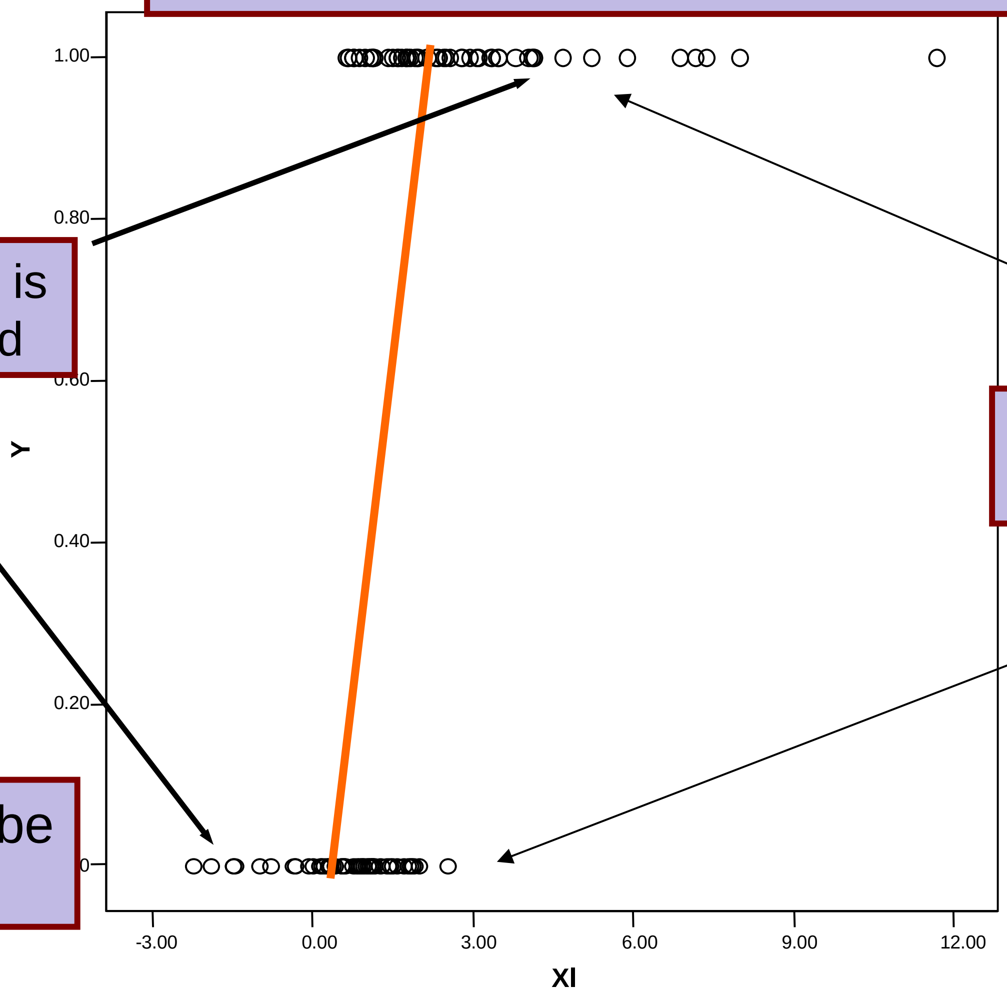
# Example




If we fit a linear regression in the scatterplot

# GLM on Dichotomous DV

The effect can never been linear




Much of the variance is always not explained

Scores will be always along two flat lines

The error term will be always big

# GLM on Dichotomous DV

- Two important facts are known in advance about the GLM in the case the DV is dichotomous one:
  - The distribution of residuals will never be normal
  - The model will produce predicted values (so fitted values) that do not make sense

# Residuals of the regression

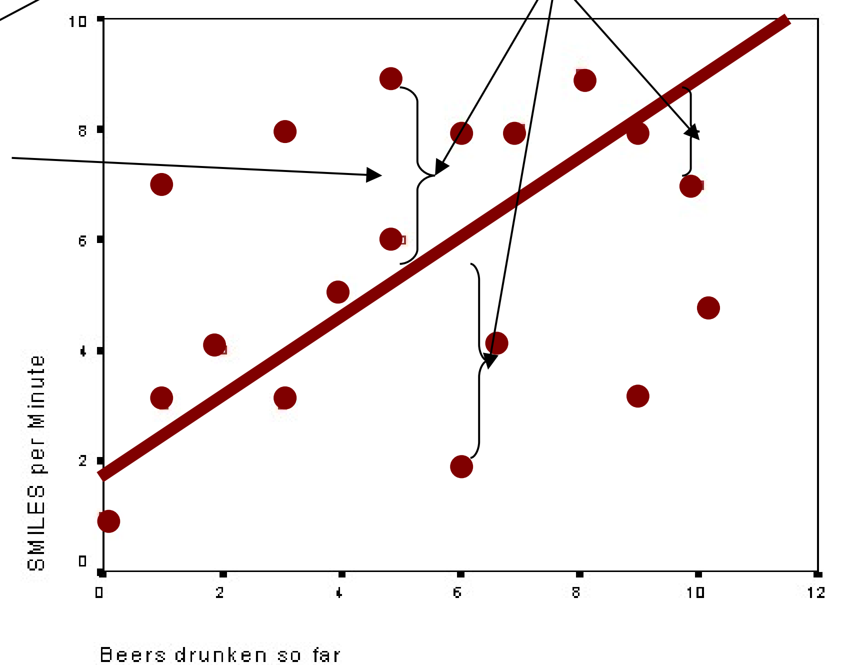
- Recall that the errors of the regressions are the differences between the predicted values and the observed values of the DV

$$\hat{y}_i = a + b_{yx} x_i$$

Predicted

Differences

Errors

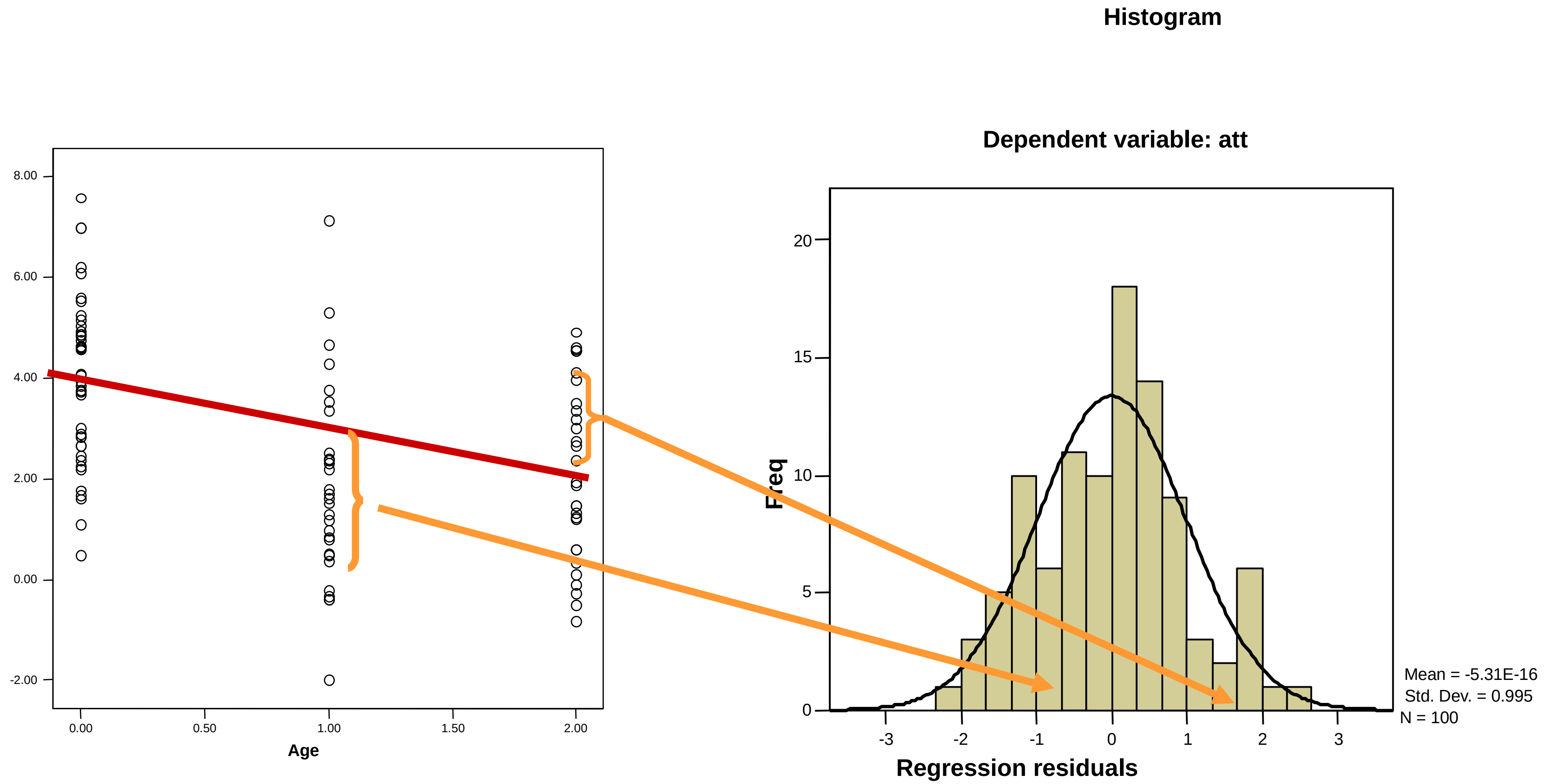



$$e_i = y_i - \hat{y}_i = y_i - (a + b_{yx} x_i)$$

Residuals: error of the regression in predicting participant’s value

# Errors Distribution

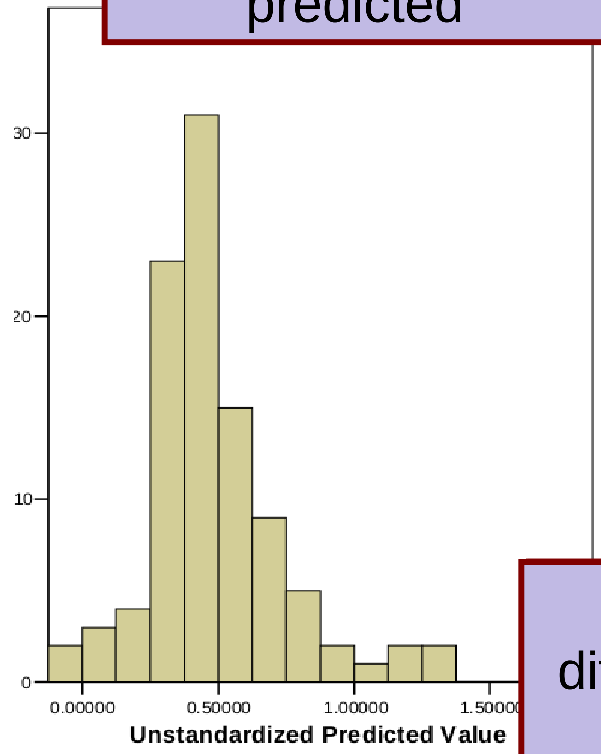
- In GLM, the assumption is that these errors are normally distributed, meaning that their frequency histogram is bell-shaped



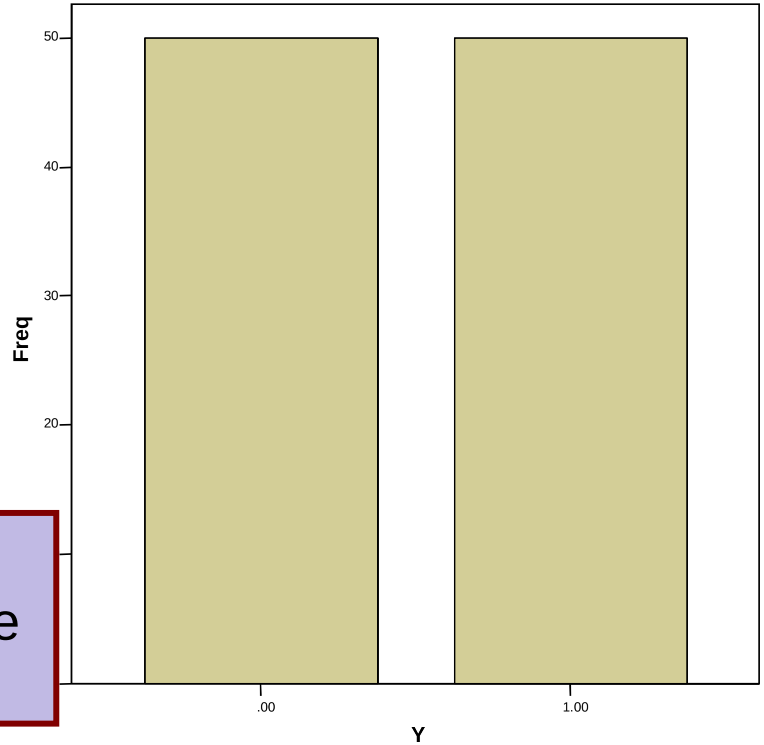
# Predicted and Observed Values

- When the DV is dichotomous, the predicted values varies across the range of possible values, the observed only assume 1 or 0 values

predicted




Observed



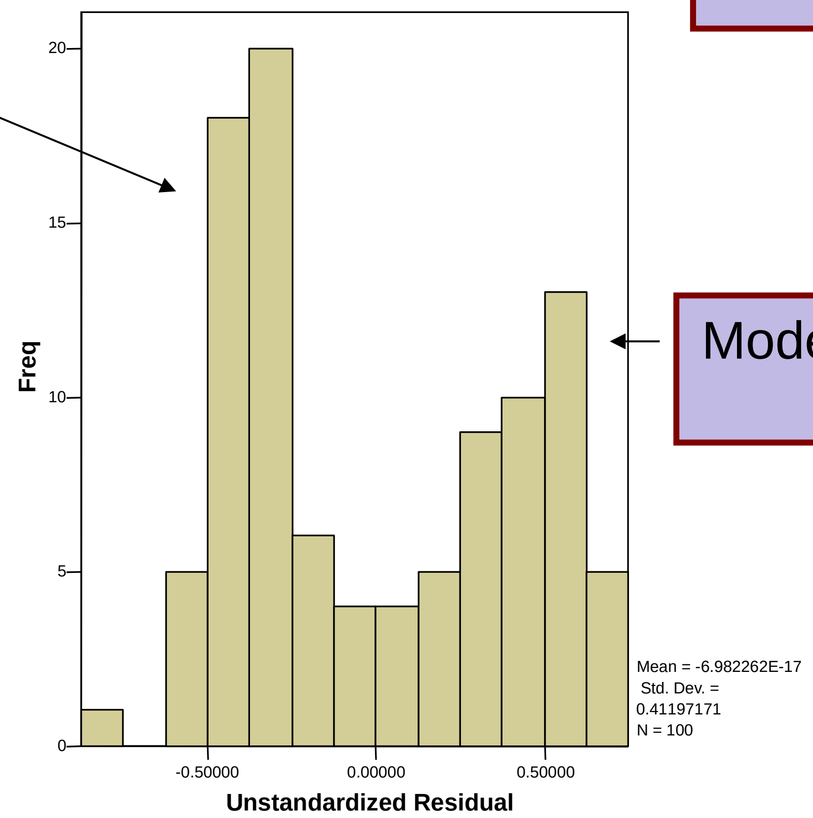

Residuals are differences between the two

# Violation of Normality

- The residuals will always have a bimodal distribution

Mode for observed value 0

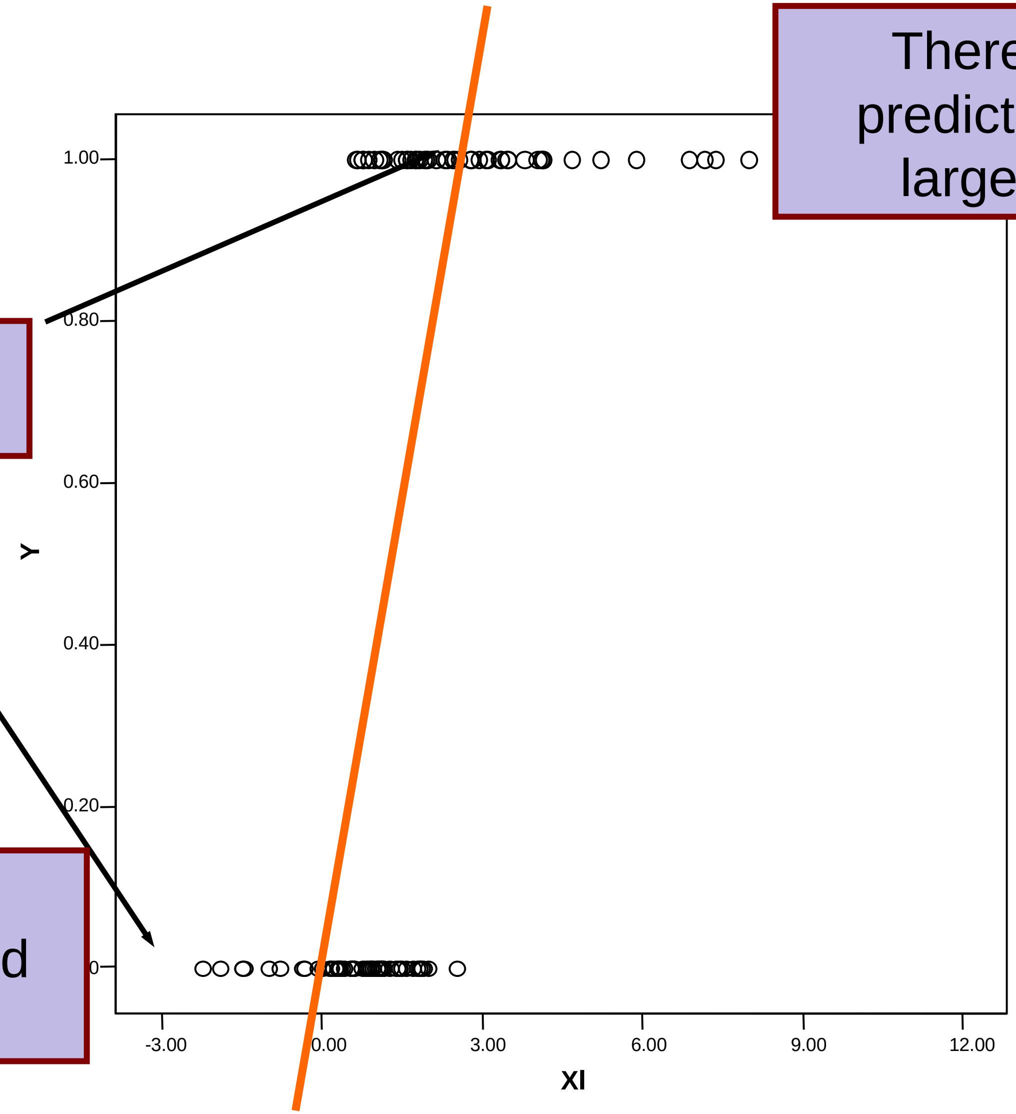
Residuals




Mode for observed value 1

# GLM on Dichotomous DV

- When the DV is dichotomous, the predicted values will not make sense




There will be predicted values larger than 1

The line will predict absurd values

There will be negative predicted values

# Dichotomous DV

- As we have seen, a variable is a dichotomy if each participant has either 1 or 0 as values.

- There are billions of them, but in psychology dichotomous DV are often "choice behavior"

- The average of the DV is the probability of observing the value 1

$$\bar{Y} = \frac{n_1}{n_{tot}}$$

- Thus, when we want to predict a dichotomous variable, we are predicting the probability of observing the value 1 (or belonging to the group with DV=1)

# Solution

## Logistic regression

$$f(y_i) = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \,..\, b_k \cdot x_{ki} + e_i$$

- Find a link function (transforms the dependent variable) to:
  - Overcome the boundaries of 1 and 0
  - Linearize the relationship
  - Obtain sensible predicted values

# Solution

## Logistic regression

$$f(y_i) = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + .. b_k \cdot x_{ki} + e_i$$

Find a residual distribution (assume a different distribution) to fit the actual distribution of the DV

# Solution: part 1

- First, instead of trying to predict the probability, we try to predict the **odds** of being in one group (DV=1) rather than the other (DV=0)

- If we try to predict if somebody would choose an option or not, we now predict the odds of choosing the option on not choosing the option

- **odd: Probability of 1 over the probability of 0**

$$P_i = a + b_{yx} x_i \quad \Rightarrow \quad \frac{P_i}{1 - P_i} = a + b_{yx} x_i$$

# Odds properties

- The odds transformation makes the dependent variable unbounded in the positive range (it varies from 0 to infinity)

$$Odd_i = \frac{P_i}{1 - P_i}$$

- **Example: if the probability of having a daughter is .50**

$$Odd = \frac{.5}{1 - .5} = 1$$

- **If the probability of voting democrats is .70**

$$Odd = \frac{.7}{1 - .7} = 2.33$$

# Odds

- Odds indicate how many times is more likely the value 1 over the value 0

$$Odd = \frac{P_i}{1 - P_i}$$

$$Odd = \frac{.5}{1 - .5} = 1$$

- **Example: A daughter is as likely as a son**

**Example: Voting democrats is 2.33 times more likely than not voting them**

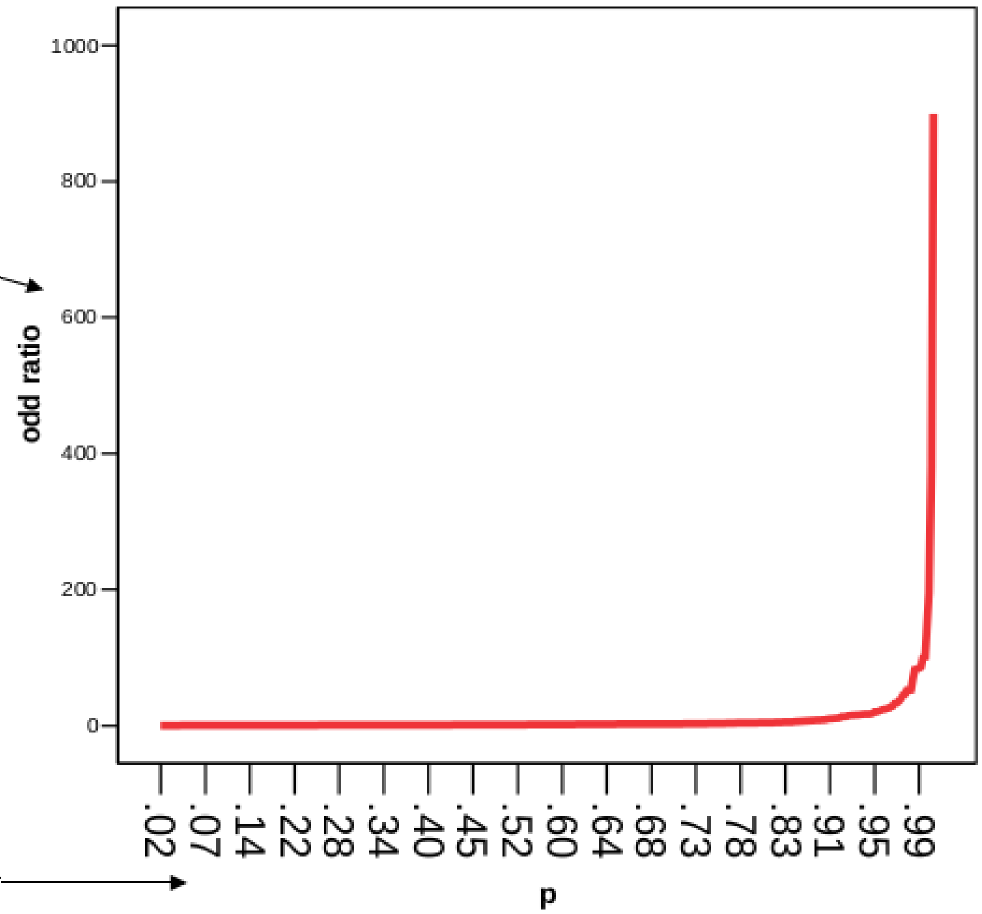
$$Odd = \frac{.7}{1 - .7} = 2.33$$

# Odds range

- The odds carry the same information than the probability, but as a variable that range from 0 to infinity

Odd from 0 to infinity

$$Odd = \frac{P_i}{1 - P_i}$$




Probability from 0 to 1

# Odds: Interpretation

Events are equally likely

$$p=.5 \rightarrow odd=\frac{.5}{1-.5}=1$$

The odd is greater than 1 if the event DV=1 is more likely then the DV=0

$$p=.7 \rightarrow odd=\frac{.7}{1-.7}=2.33>1$$

The odd is less than 1 if the event DV=1 is less likely then the DV=0

$$p=.2 \rightarrow odd=\frac{.2}{1-.2}=.25<1$$

# The problem with odds

- If we try to use the GLM machinery to predict odds, however, we will have negative prediction that do not make sense *a priori*

$$\frac{P_i}{1-P_i} = a + b_{yx} x_i$$

If a=1, b=3 e x=-2

$$\frac{P_i}{1-P_i} = 1 + 3*(-2) = -5$$

# Solution: part 2

- Instead of predicting the odds, we predict the logarithm of the odds

$$\frac{P_i}{1-P_i} = a + b_{yx}x_i \Rightarrow \boxed{\ln\left(\frac{P_i}{1-P_i}\right) = a + b_{yx}x_i}$$

The logarithm transformation is called **logit**

$$\log it = \ln\left(\frac{P_i}{1-P_i}\right)$$

The regression that predict the logit is called the logistic regression

# Logarithm

- Exponent of the power to which it is necessary to raise a fixed number (the base) to produce the given number. For example, the logarithm of 100 (base 10) is 2 because $10^2$ equals 100.

$$Log_{10}(100) = 2$$

- We often use the (*napierian*) natural logarithm, which is the power to which it is necessary to raise *e* to obtain a given number

$$e = 2.71828182845904523536028747135266249775 7\ldots$$

$$e^{4.605} = 100 \qquad Ln(100) = 4.605$$

# Why the Logarithm

- The logistic uses the logarithm because:
    - Transforms the odds in negative and positive
    - Is positive if the odd is greater than 1
    - Is negative if the odd is less than 1
    - Is zero if the odd is 1

# Logit Transformation

- Basically, we transform the probability such that it can assume values that make sense when predicted with a regression

$$P_i(Y=1)$$
$$1-P_i(Y=0)$$

$$Odd = \frac{P_i}{1-P_i}$$

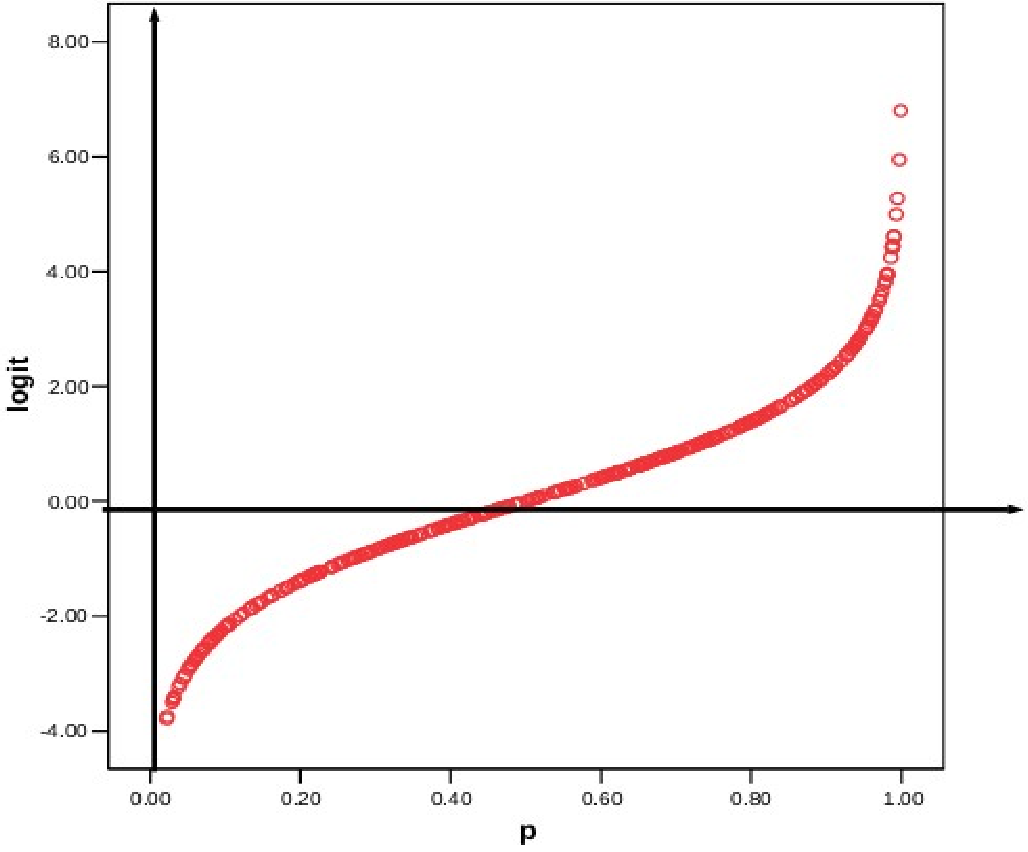
$$\ln\left(\frac{P_i}{1-P_i}\right)$$

| Said YES | How more likely is to say YES over say NO | A continuous variable across negative and positive values |
|---|---|---|
| P=.80 | Odd=4 | Ln=1.38 |
| P=0.50 | Odd=1 | Ln=0 |
| P=.20 | Odd=.25 | Ln=-1.38 |

# The Logit

- All possible predictions make sense, because the logit varies from negative to positive

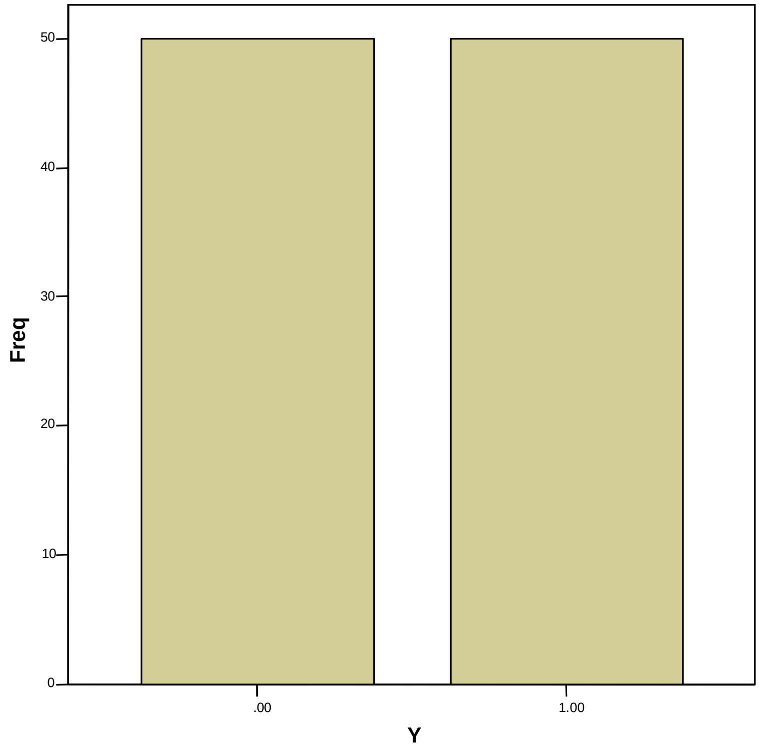
$$\log it = \ln\left(\frac{p}{1-p}\right)$$



# Assumption about the distribution

- In logistic regression, the assumption is that the data come from a binomial distribution, in which values only assume 1 or 0 values

Binomial Distribution



# A case of GzLM

$$\ln\left(\frac{P_i}{1-P_i}\right) = a + b_{yx}x_i$$

- If the logistic is simply a regression on the logit, the logic of regression can be used in the logistic (that is nice)

- As for any regression, the B coefficients are expressed in terms of the scale of the dependent variable (this is not nice)

# Example

- I want to establish if there is an effect of extroversion on people preference for beer or wine

$$\ln\left(\frac{Beer}{Wine}\right) = a + b_{yx}\, EXTRO_i$$

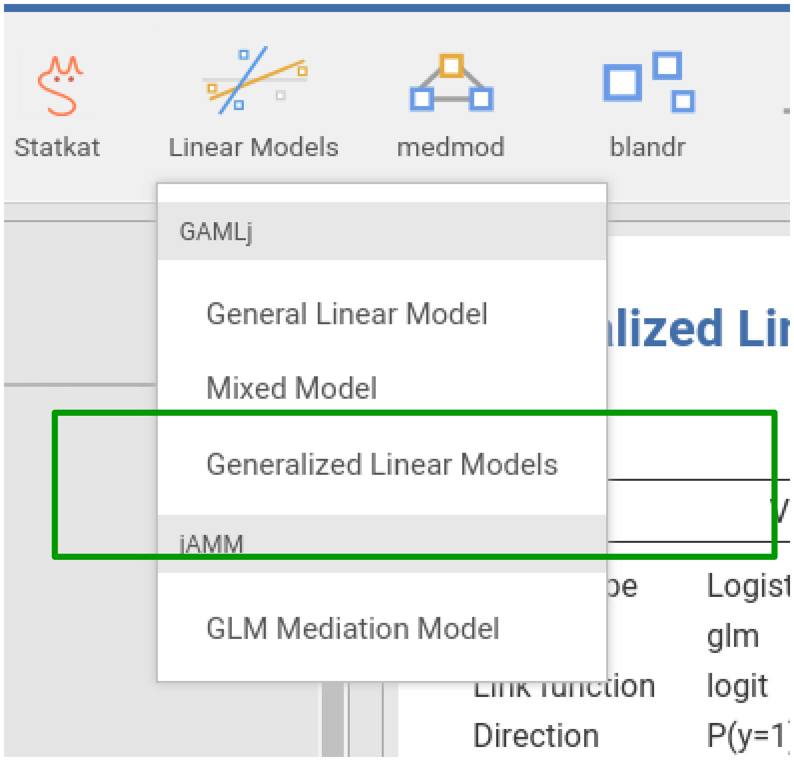
# jamovi Data

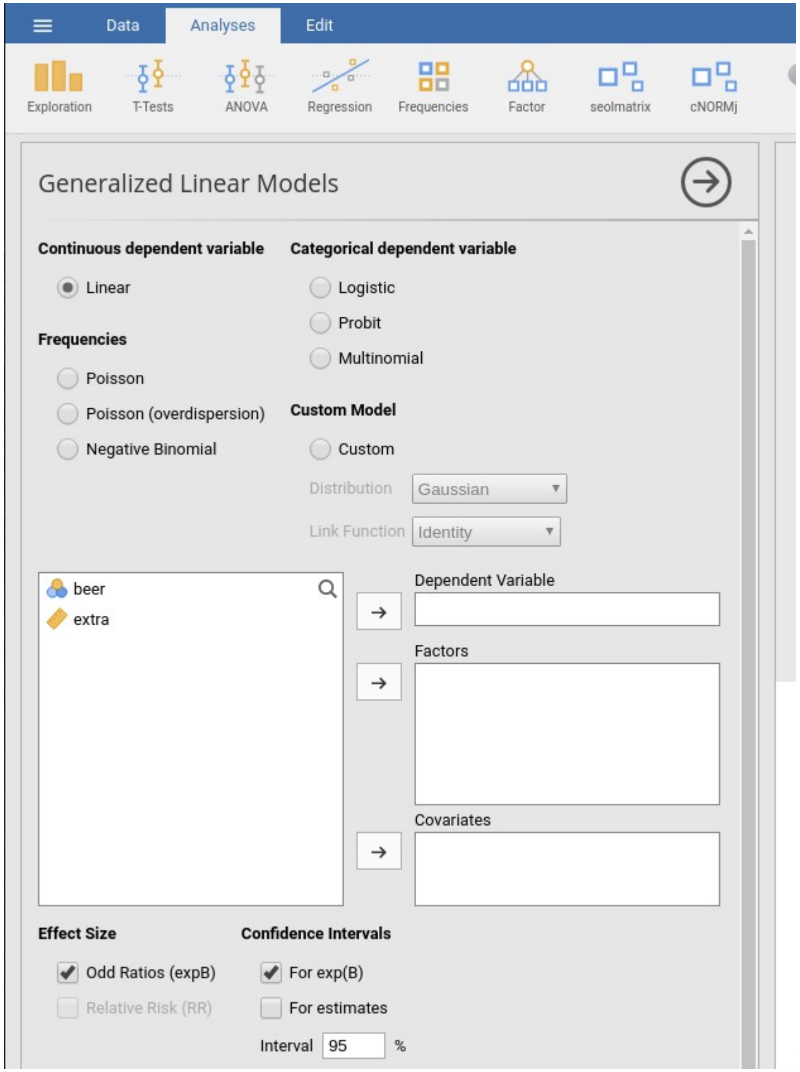
| | beer | extra |
|---|---|---|
| 1 | 0 | 12.97 |
| 2 | 0 | 11.29 |
| 3 | 0 | 9.55 |
| 4 | 0 | 9.48 |
| 5 | 1 | 12.02 |
| 6 | 1 | 9.61 |
| 7 | 0 | 10.24 |
| 8 | 1 | 11.65 |
| 9 | 1 | 11.10 |
| 10 | 0 | 10.50 |
| 11 | 0 | 9.72 |
| 12 | 1 | 11.15 |
| 13 | 1 | 10.28 |
| 14 | 0 | 9.52 |
| 15 | 1 | 12.86 |
| 16 | 0 | 9.88 |
| 17 | 1 | 11.92 |
| 18 | 0 | 8.61 |
| 19 | 0 | 7.95 |
| 20 | 1 | 11.75 |
| 21 | 1 | 11.55 |
| 22 | 0 | 9.17 |
| 23 | 1 | 11.06 |

**Dichotomous Dependent variable**

# jamovi GzLM

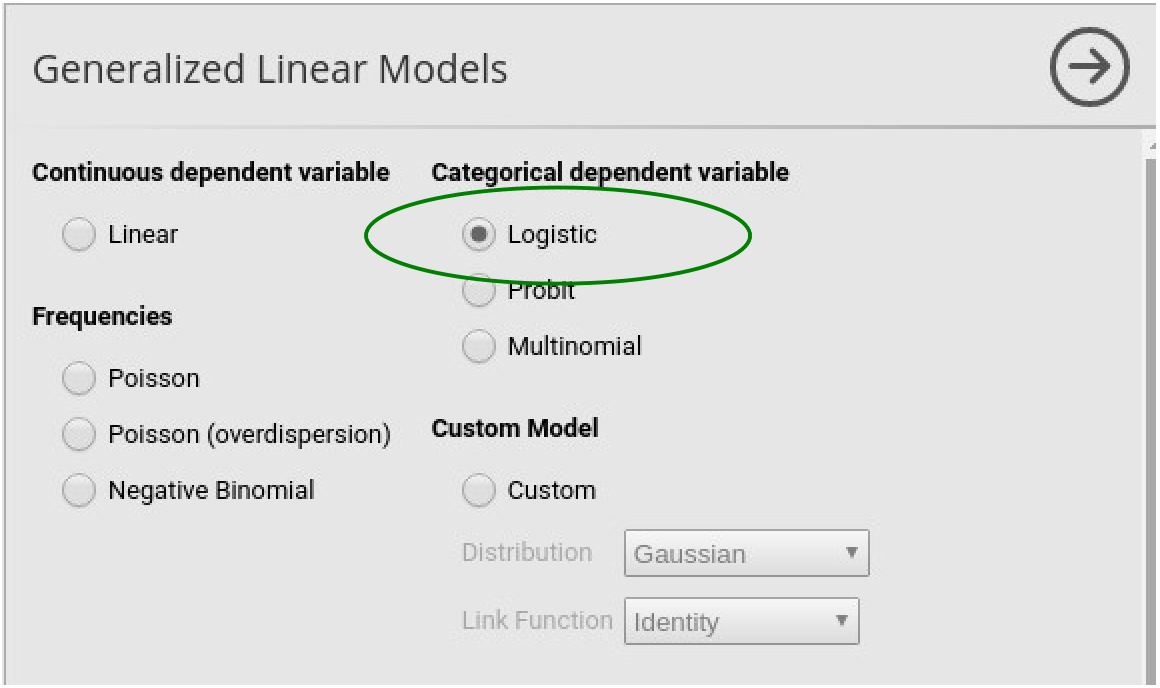
- We use "generalized linear models", which can be used for several different types of GzLM models



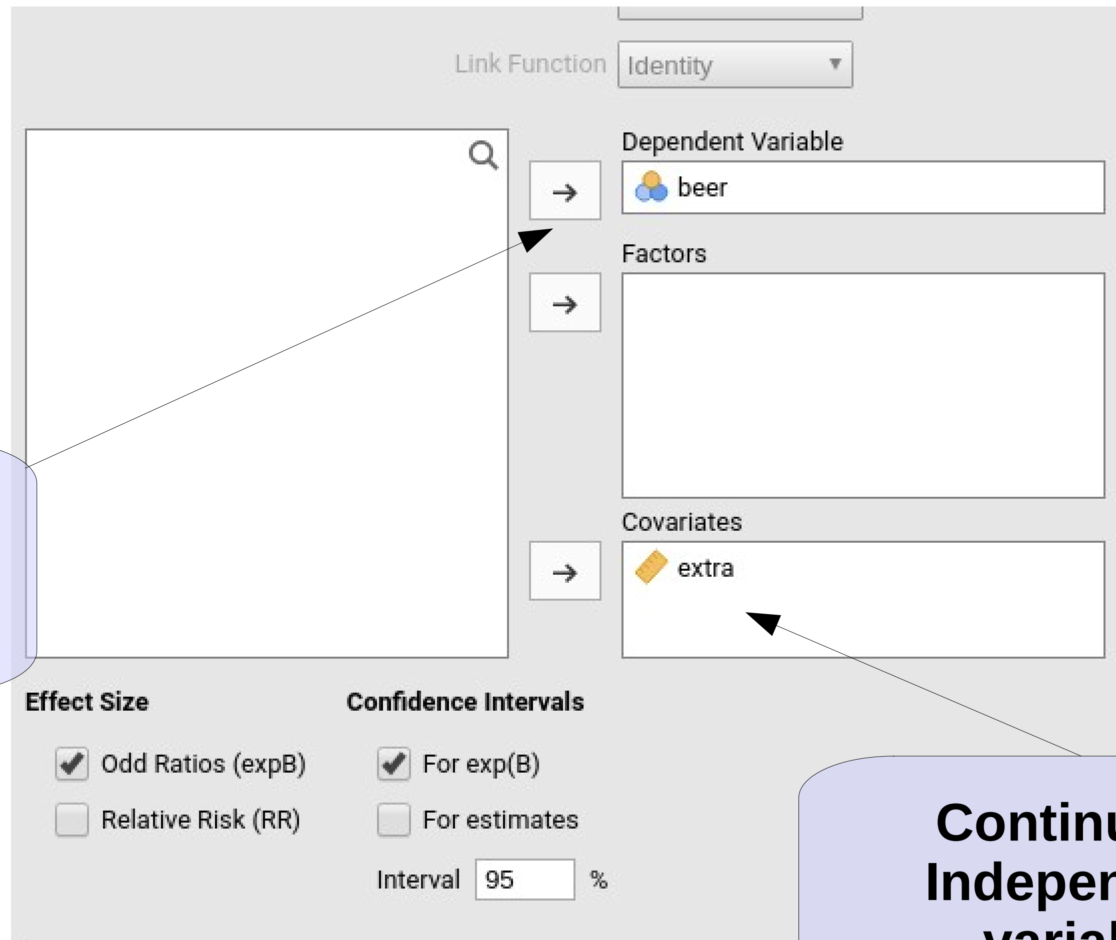
# jamovi

- First, we select the type of model we need: logistic



# jamovi

- We set the roles of the variables




**Dichotomous Dependent variable**

**Continuous Independent variable**

# Results

- Info about the model

Model

## Generalized Linear Models

Model Info

| Info | Value | Comment |
|---|---|---|
| Model Type | Logistic | Model for binary y |
| Call | glm | beer ~ 1 + extra |
| Link function | Logit | Log of the odd of y=1 over y=0 |
| Direction | P(y=1)/P(y=0) | P( beer = 1 ) / P( beer = 0 ) |
| Distribution | Binomial | Dichotomous event distribution of y |
| R-squared | 0.245 | Proportion of reduction of error |
| AIC | 54.831 | Less is better |
| Deviance | 50.831 | Less is better |
| Residual DF | 48 | |
| Chi-squared/DF | 1.050 | Overdispersion indicator |
| Converged | yes | Whether the estimation found a solution |

[3]

# Coefficients

- Coefficients should be interpret as in regression: The expected change when you move the IV of one unit

Parameter Estimates

| | | | | 95% Exp(B) Confidence Interval | | | |
|---|---|---|---|---|---|---|---|
| Names | Estimate | SE | exp(B) | Lower | Upper | z | p |
| (Intercept) | −0.536 | 0.350 | 0.585 | 0.283 | 1.14 | −1.53 | 0.126 |
| extra | 1.146 | 0.342 | 3.145 | 1.722 | 6.71 | 3.35 | < .001 |

$$\ln\left(\frac{beer}{wine}\right) = -0.536 + 1.146\, EXTRO_i$$

# Constant term

- Coefficients should be interpret as in regression

- The expected value of the DV for the IV equal to zero

$$\ln(odd_0) = a + b_{yx} 0$$

$$\ln(odd_0) = a$$

# Coefficient B

- Coefficients should be interpret as in regression

- The expected change when you move the IV of one unit

$$\ln(odd_0) = a + b_{yx} 0$$

$$\ln(odd_1) = a + b_{yx} 1$$

$$b_{yx} = \ln(odd_1) - \ln(odd_0)$$

# Coefficients

- Coefficients should be interpret as in regression

- Thus they tell us the change in logarithm of the odds (?!)

Parameter Estimates

| | | | | 95% Exp(B) Confidence Interval | | | |
|---|---|---|---|---|---|---|---|
| Names | Estimate | SE | exp(B) | Lower | Upper | z | p |
| (Intercept) | −0.536 | 0.350 | 0.585 | 0.283 | 1.14 | −1.53 | 0.126 |
| extra | 1.146 | 0.342 | 3.145 | 1.722 | 6.71 | 3.35 | < .001 |

Differences in the log of the odds

$$b_{yx} = \ln(odd_{x+1}) - \ln(odd_{x})$$

# EXP(B)

- To interpret the results, we remove the logarithm scale from the B by applying the exponential transformation

$$\exp(\ln(x)) = x$$

- If you take the exponential transformation on a logarithm, the result will be expressed in the scale (units) of the argument of the log

$$\exp(\ln(meters)) = meters$$

# EXP(B) constant

- By removing the log scale, we obtain the odds for X=0

$$a = \ln(odd_0)$$

$$\exp(a) = odd_0$$

- How more likely is the DV=1 for the independent variable equal to zero

- How more likely is to choose wine rather than beer for 0 extroversion

# Coefficients

- Coefficients should be interpret as in regression

- EXP constant coefficient is expressed in odds scale

Parameter Estimates

| | | | | 95% Exp(B) Confidence Interval | | | |
|---|---|---|---|---|---|---|---|
| Names | Estimate | SE | exp(B) | Lower | Upper | z | p |
| (Intercept) | −0.536 | 0.350 | 0.585 | 0.283 | 1.14 | −1.53 | 0.126 |
| extra | 1.146 | 0.342 | 3.145 | 1.722 | 6.71 | 3.35 | < .001 |

For extroversion = 0, preferring beer is 0.585 times more likely than choosing wine

# EXP(B): Slope

- The only difficulty to overcome is remember that the exponential of a log difference is equal to a ratio

$$\ln(a) - \ln(b) = q$$

$$\exp(\ln(a) - \ln(b)) = \exp(q)$$

$$\exp(q) = \frac{a}{b}$$

# EXP(B): Odd ratio

- Thus the exp(B) is the odd ratio between two consecutive odds, as you move the IV of 1 unit

$$b_{yx} = \ln(odd_{x+1}) - \ln(odd_x)$$

$$\exp(b_{yx}) = \frac{odd_{x+1}}{odd_x}$$

- Thus the exp(B) is **how many times** the odd changes as you move the independent variable of one unit

# Coefficients: Effects

- The odd ratio exp(B) tells us how many times the odd of wine over beer changes as you change the IV of one unit

Parameter Estimates

| | | | | 95% Exp(B) Confidence Interval | | | |
|---|---|---|---|---|---|---|---|
| Names | Estimate | SE | exp(B) | Lower | Upper | z | p |
| (Intercept) | −0.536 | 0.350 | 0.585 | 0.283 | 1.14 | −1.53 | 0.126 |
| extra | 1.146 | 0.342 | 3.145 | 1.722 | 6.71 | 3.35 | < .001 |

As extroversion increases of 1 unit, the odd of preferring beer over wine increases 3.145 times

# EXP(B): Odd ratio

- The odd ratio exp(B) tells us how many times the odd of wine over beer changes as you change the IV of one unit

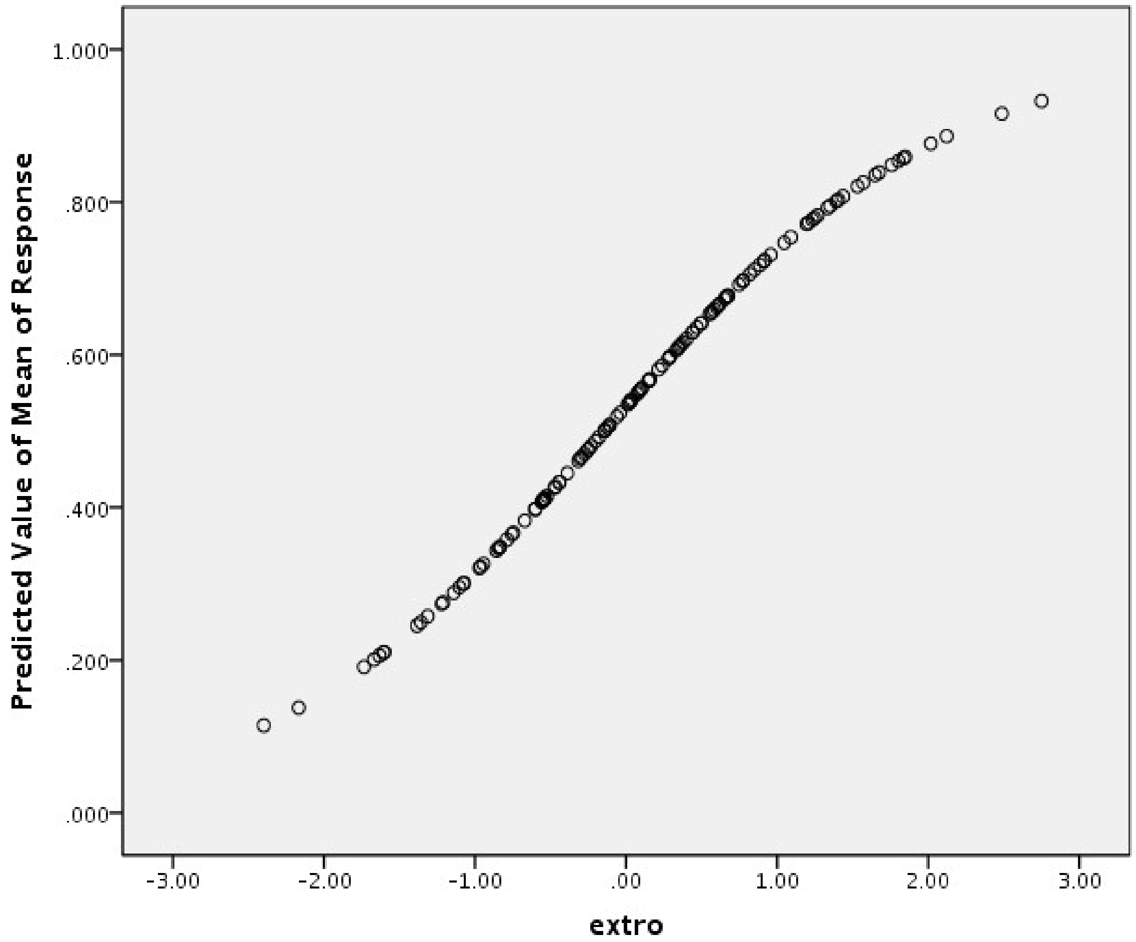
$$\exp(b_{yx}) = \frac{odd_{x+1}}{odd_x}$$

$$odd_{x+1} = \exp(b_{yx}) * odd_x$$

As extroversion increases of 1 unit, the odd of preferring beer over wine increases of 3.145 times

# Visualizing the effects

- As in regression one looks at the line, here one looks at the predicted probability of being in a group (rather than the other)



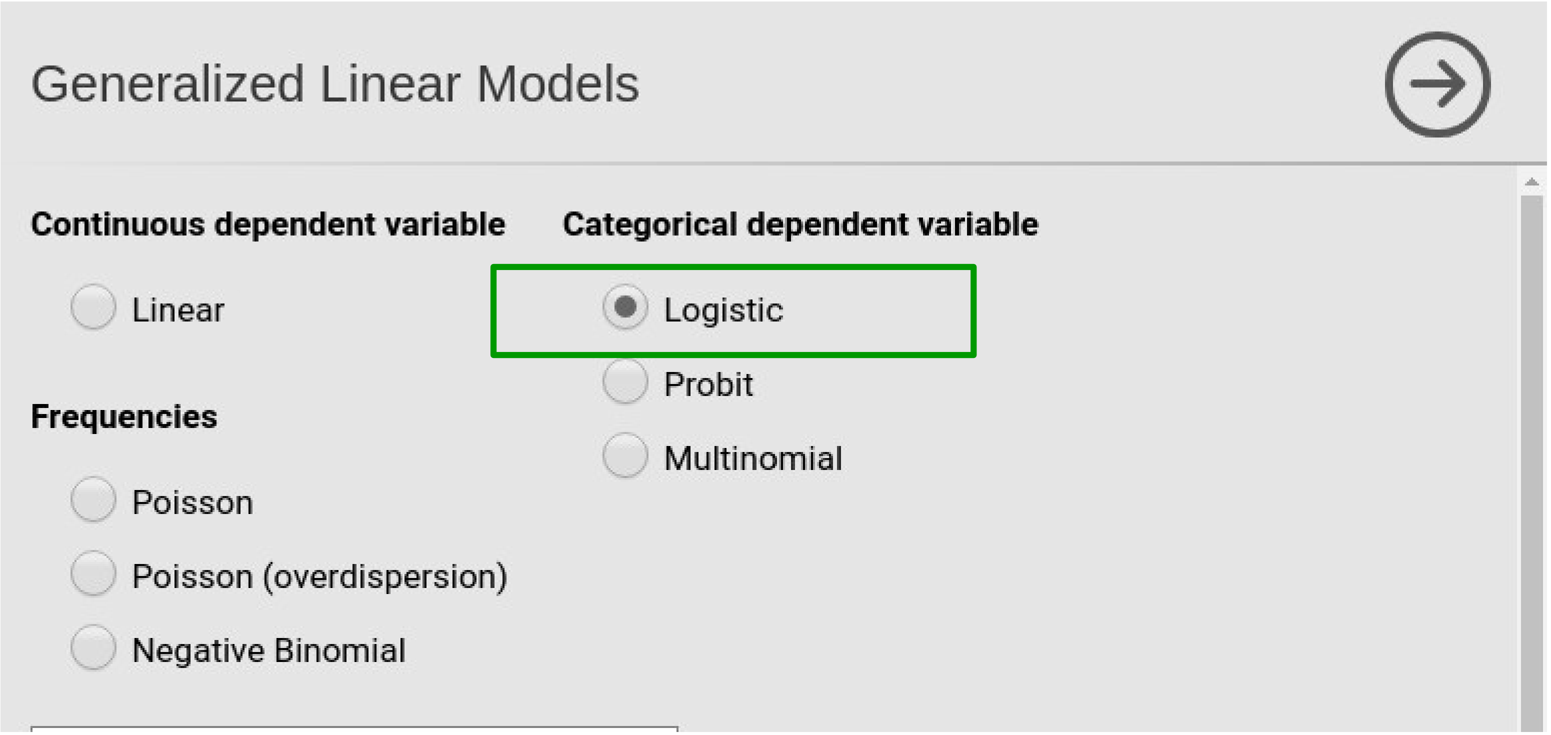
# Jamovi: example

- The data set Neuralgia contains five variables: Treatment, Sex, Age, Duration, and Pain. The last variable, Pain, is the response variable. A specification of Pain=Yes indicates there was pain, and Pain=No indicates no pain. The variable Treatment is a categorical variable with three levels: A and B represent the two test treatments, and P represents the placebo treatment. The gender of the patients is given by the categorical variable Sex. The variable Age is the age of the patients, in years, when treatment began. The duration of complaint, in months, before the treatment began is given by the variable Duration.

- Let's first consider the relationship between **pain** and **age**

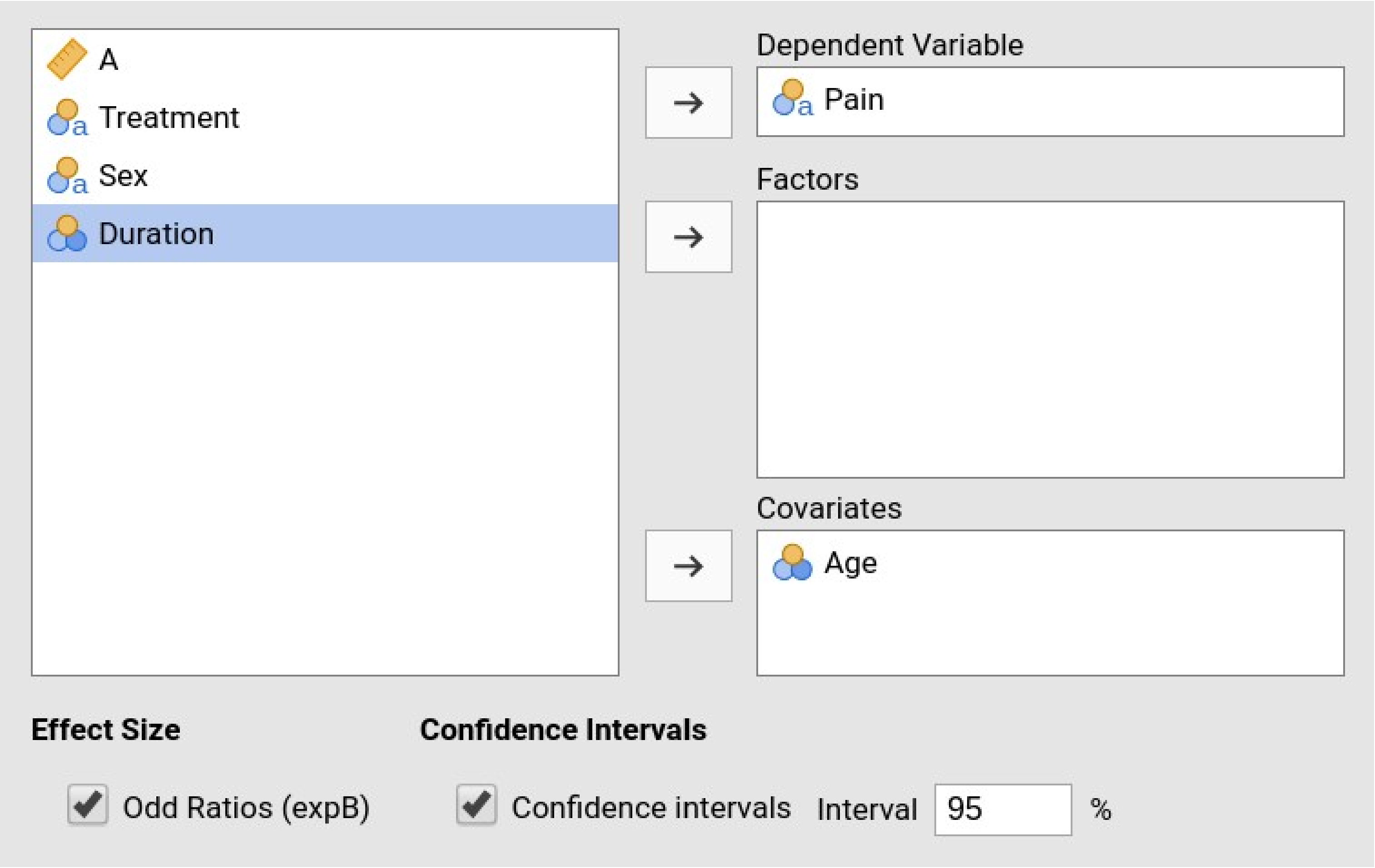
# Jamovi: example

- First we need to select the type of model we need



# Jamovi: example

- Then we define the variables role



# Jamovi: example

- Results: recap table

Model Info

| Info | Value | Comment |
|---|---|---|
| Model Type | Logistic | Model for binary y |
| Call | glm | Pain ~ 1 + Age |
| Link function | logit | Log of the odd of y=1 over y=0 |
| Direction | P(y=1)/P(y=0) | P( Pain = Yes ) / P( Pain = No ) |
| Distribution | Binomial | Dichotomous event distribution of y |
| R-squared | 0.104 | Proportion of reduction of error |
| AIC | 77.056 | Less is better |
| Deviance | 73.056 | Less is better |
| Residual DF | 58 | |
| Converged | yes | A solution was found |

The R-squared can be interpreted as in the GLM: the proportion of reduce error: how well the model fits the data

# Jamovi: example

- Results: omnibus tests and coefficients

## Model Results

Loglikelihood ratio tests

| | $X^2$ | df | p |
|---|---|---|---|
| Age | 8.45 | 1 | 0.004 |

This is equivalent to the GLM F-test

Fixed Effects Parameter Estimates

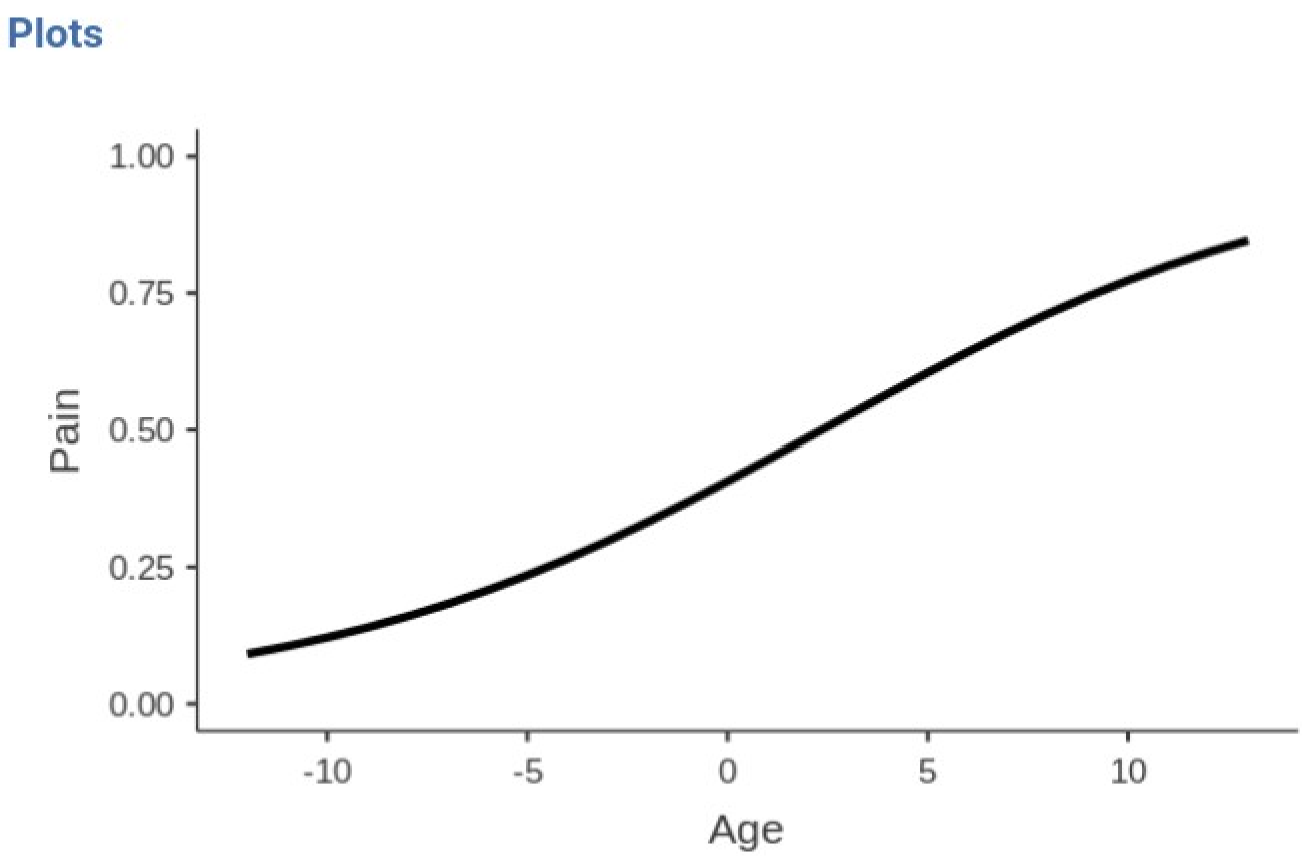
| | | | | 95% Confidence Interval | | | | |
|---|---|---|---|---|---|---|---|---|
| Names | Effect | Estimate | SE | Lower | Upper | exp(B) | z | p |
| (Intercept) | (Intercept) | −0.377 | 0.2826 | −0.9468 | 0.171 | 0.686 | −1.33 | 0.183 |
| Age | Age | 0.160 | 0.0600 | 0.0499 | 0.288 | 1.174 | 2.67 | 0.007 |

This is equivalent to the GLM B and t-test table

SPSS uses wald-test. Jamovi uses z-tests, results are equivalent

# Jamovi: example

- Results: plot of probabilities (of being in group 1)




With increasing age, the prob. of feeling pain increases

# Logistic on ANOVA designs

- The data set Neuralgia contains five variables: Treatment, Sex, Age, Duration, and Pain. The last variable, Pain, is the response variable. A specification of Pain=Yes indicates there was pain, and Pain=No indicates no pain. The variable Treatment is a categorical variable with three levels: A and B represent the two test treatments, and P represents the placebo treatment. The gender of the patients is given by the categorical variable Sex. The variable Age is the age of the patients, in years, when treatment began. The duration of complaint, in months, before the treatment began is given by the variable Duration.

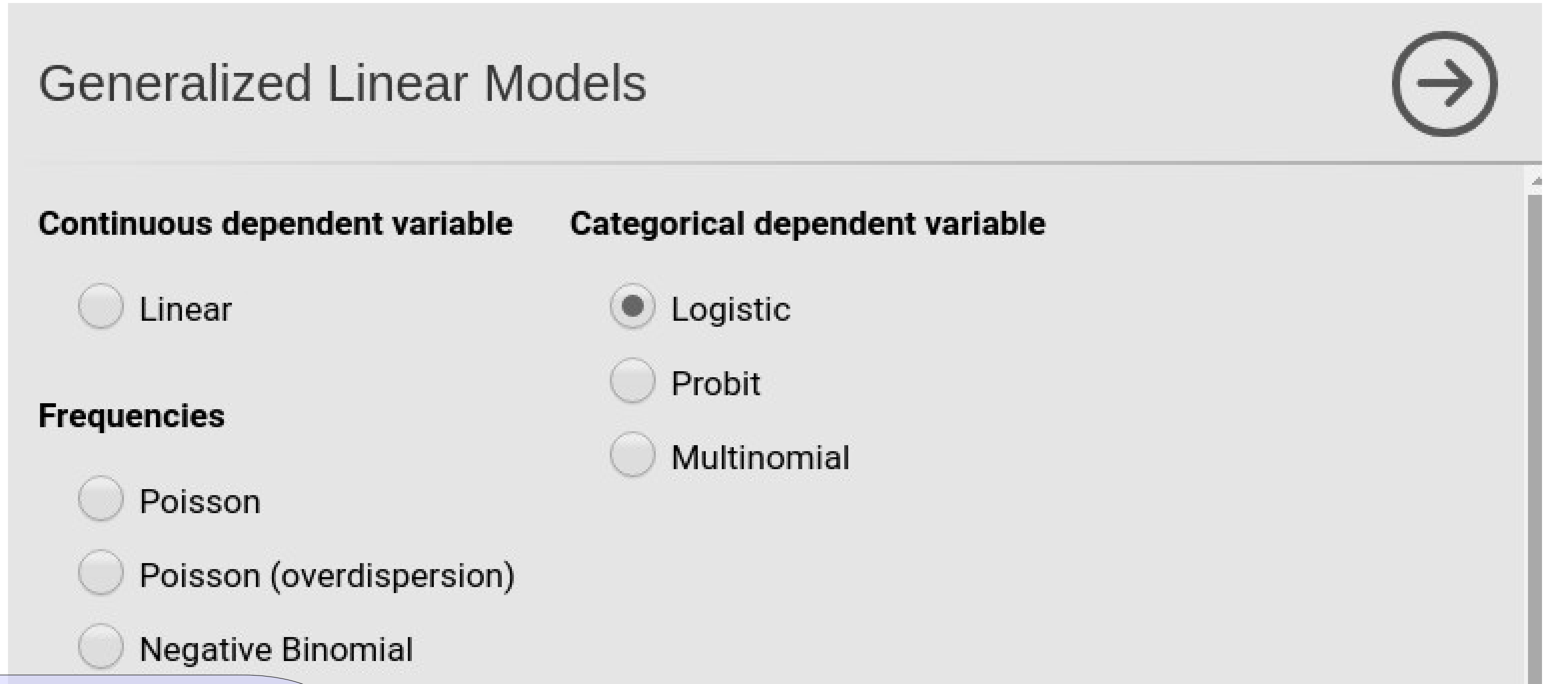
- Let's first consider the relationship between **pain** and **sex and treatment**

# Logistic on ANOVA designs

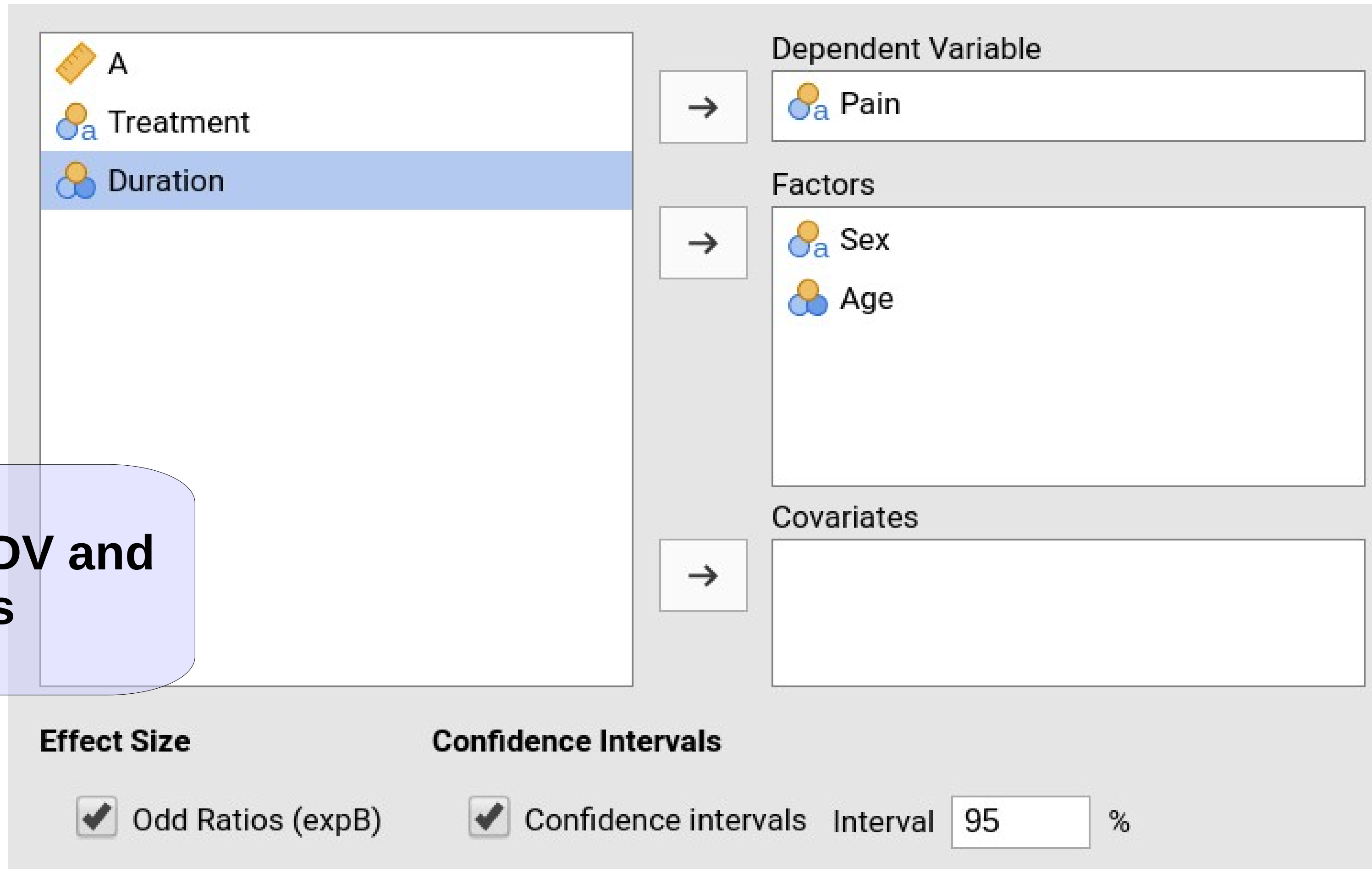
- Let's first consider the relationship between **pain** and **sex and treatment**

- We simply apply the logic of the ANOVA (as in GLM) knowing that the predicted values are the *logit of the odd of being in group 1 rather than group 0*

# jamovi




**Specify the type of model**




**Specify the DV and factors**

# Coefficients

- Coefficients are expressed in the logarithmic scale (B) and in the odd ratios scale exp(B)

**Model Results**

Main effects and interactions (as the F in GLM)

Loglikelihood ratio tests

| | $X^2$ | df | p |
|---|---|---|---|
| Sex | 7.59 | 1 | 0.006 |
| Treatment | 15.96 | 2 | < .001 |
| Sex ✻ Treatment | 7.11e−15 | 2 | 1.000 |

B coefficients and Odd ratios B

Fixed Effects Parameter Estimates

| | | | | 95% Confidence Interval | | | | |
|---|---|---|---|---|---|---|---|---|
| Names | Effect | Estimate | SE | Lower | Upper | exp(B) | z | p |
| (Intercept) | (Intercept) | −0.434 | 0.357 | −1.191 | 0.2786 | 0.648 | −1.22 | 0.224 |
| Sex1 | M - ( F, M ) | 0.896 | 0.357 | 0.246 | 1.6859 | 2.449 | 2.51 | 0.012 |
| Treatment1 | B - ( A, B, P ) | −0.868 | 0.505 | −2.025 | 0.0694 | 0.420 | −1.72 | 0.086 |
| Treatment2 | P - ( A, B, P ) | 1.735 | 0.505 | 0.834 | 2.9049 | 5.670 | 3.44 | < .001 |
| Sex1 ✻ Treatment1 | M - ( F, M ) ✻ B - ( A, B, P ) | −4.92e−16 | 0.505 | −0.972 | 1.1430 | 1.000 | −9.75e−16 | 1.000 |
| Sex1 ✻ Treatment2 | M - ( F, M ) ✻ P - ( A, B, P ) | 6.44e−16 | 0.505 | −0.972 | 1.1430 | 1.000 | 1.28e−15 | 1.000 |

# Visualizing the effect

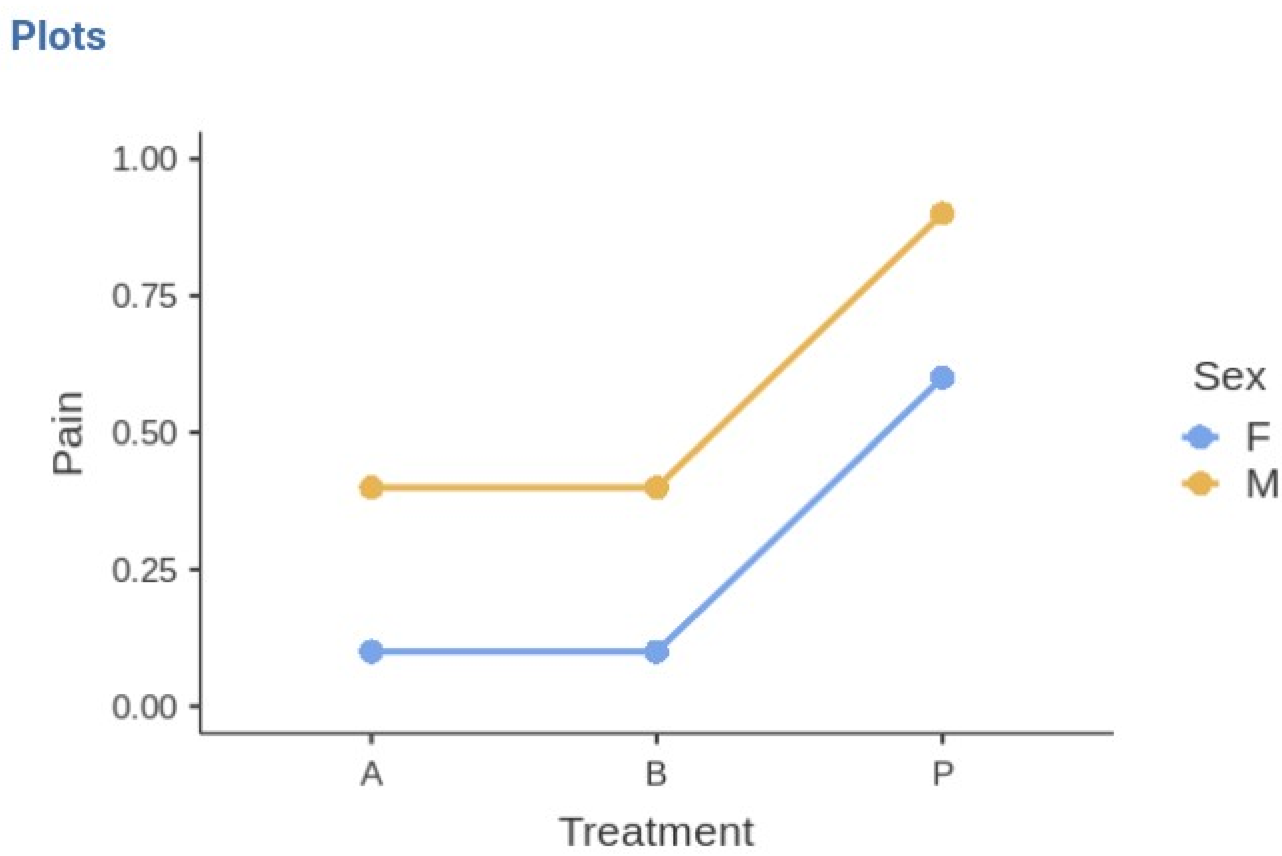
- As in ANOVA one looks at the cell means, here one looks at table of probabilities of being in group 1 (pain=yes)

Sex:Treatment

| | | | | 95% Confidence Interval | |
|---|---|---|---|---|---|
| Sex | Treatment | Prob. | SE | Lower | Upper |
| F | A | 0.100 | 0.0949 | 0.0139 | 0.467 |
| M | A | 0.400 | 0.1549 | 0.1583 | 0.703 |
| F | B | 0.100 | 0.0949 | 0.0139 | 0.467 |
| M | B | 0.400 | 0.1549 | 0.1583 | 0.703 |
| F | P | 0.600 | 0.1549 | 0.2974 | 0.842 |
| M | P | 0.900 | 0.0949 | 0.5328 | 0.986 |

# Visualizing the effect

- Or at the plots of probabilities



# Regression vs Logistic

- All we know about regression/ANOVA (interaction, partial effects, intercept, covariate, dummies for categorical IVs) remains the same for logistic models

- The difference lies in the interpretation of the coefficients

# Recap

- **Logistic regression** computes a regression with a dichotomous dependent variable
- The coefficients are expressed in the logarithmic scale (B) and as odd ratios exp(B)
- The exp(B) is the amount the odd ratio is multiplied when we move the independent variable of 1 unit
- Goodness of fit is measured with likelihood ratio, and approximation of $R^2$
- Overall significance is test with the Chi-square test

# Generalized linear model

- Applying this logic we obtain a large set of possible statistical techniques

$$f(y_i) = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + .. b_k \cdot x_{ki} + e_i$$

| Dependent Variable | function | Distribution |
| --- | --- | --- |
| Continuous | identity | Normal |
| Dichotomous | Logit of odd | Binomial |
| Categorical | Logit of odd | Multinomial |
| Ordinal | Cumulative Logit | Multinomial |
| Frequencies | Frequencies LN | Poisson |

Multinomial model

Dependent variable with more then two groups

# Theory

The Multinomial model decomposes the dependent variable in K-1 dummies, where K is the number of levels, and uses logistic regression to estimate the effects of the independent variables on these dummies

| Groups | dummy1 | dummy2 |
|---|---|---|
| A | 0 | 0 |
| B | 0 | 1 |
| C | 1 | 0 |

# Theory

The Multinomial model decomposes the dependent variable in K-1 dummies, where K is the number of levels, and uses logistic regression to estimate the effects of the independent variables on these dummies

| Groups |
|---|
| A |
| B |
| C |

dummy1

$$\ln\left(\frac{p(B)}{p(A)}\right)=a_1+b_1 X$$

dummy2

$$\ln\left(\frac{p(C)}{p(A)}\right)=a_2+b_2 X$$

# Theory

- It produces an intercept and a coefficient for each dummy
- And an omnibus effect testing that all effects are zero

If there are more IV, each IV has a coefficient for each dummy

| Groups |
|---|
| A |
| B |
| C |

dummy1

$$\ln\left(\frac{p(B)}{p(A)}\right)=a_1+b_1 X$$

dummy2

$$\ln\left(\frac{p(C)}{p(A)}\right)=a_2+b_2 X$$

# Example

The data set contains variables on 200 students. The outcome (dependent) variable is **prog**, program type. There are three programs that students can choose: general program, vocational program and academic program. The predictor (independent) variables are social economic status, **ses**, a three-level categorical variable and writing score, **write**, a continuous variable (UCLA idre web page).

We ask whether ability to **write** influences the **prog**ram choice

## Contingency Tables

Contingency Tables

| | ses | | | |
|---|---|---|---|---|
| prog | high | low | middle | Total |
| academic | 42 | 19 | 44 | 105 |
| general | 9 | 16 | 20 | 45 |
| vocation | 7 | 12 | 31 | 50 |
| Total | 58 | 47 | 95 | 200 |

hdsdemo.csv

# Example

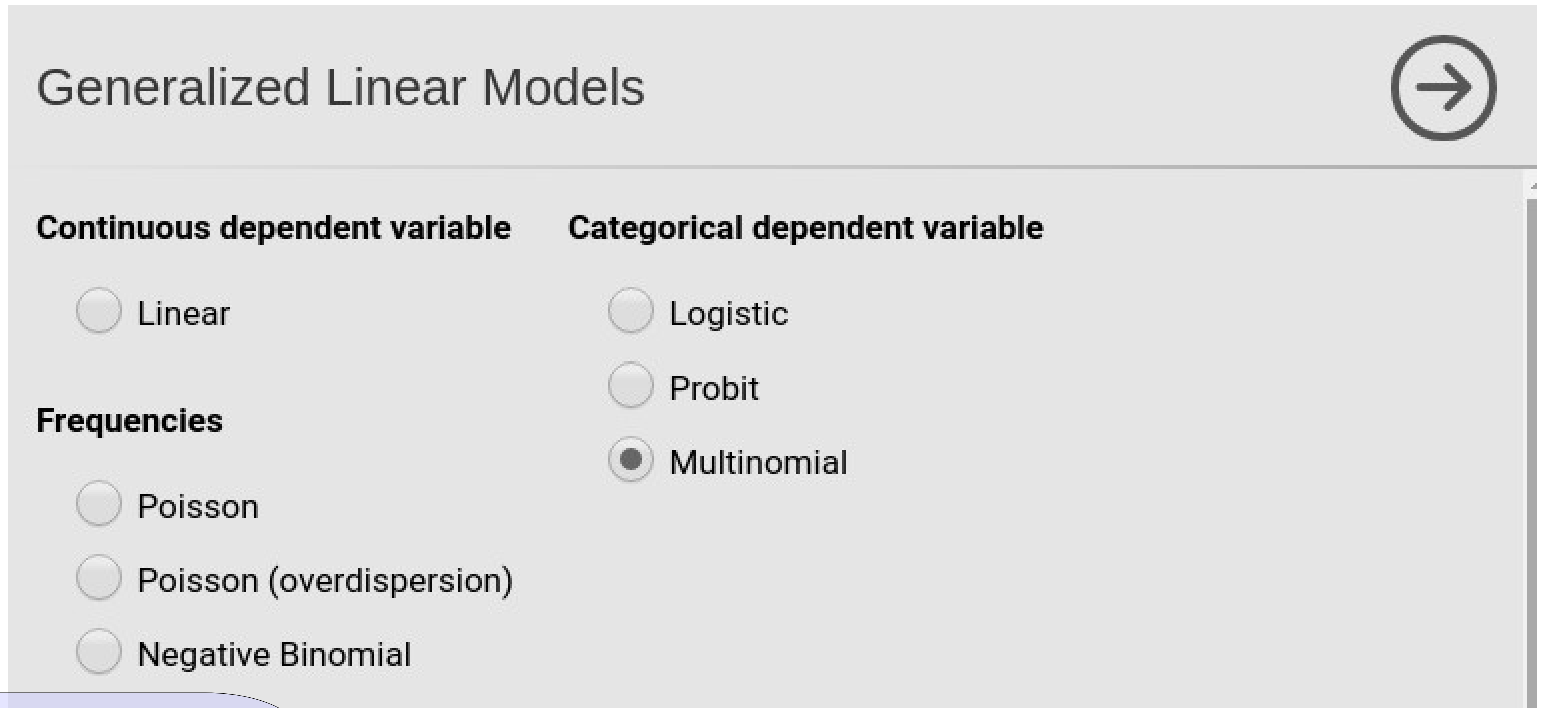
The model picks a reference group for the dependent variable, say **general program**

The model estimates the influence of the independent variable(s) on the logit (log of odd) of choosing each program over the academic program.

Having three programs, our analysis will estimate two (K-1) sets of coefficients:

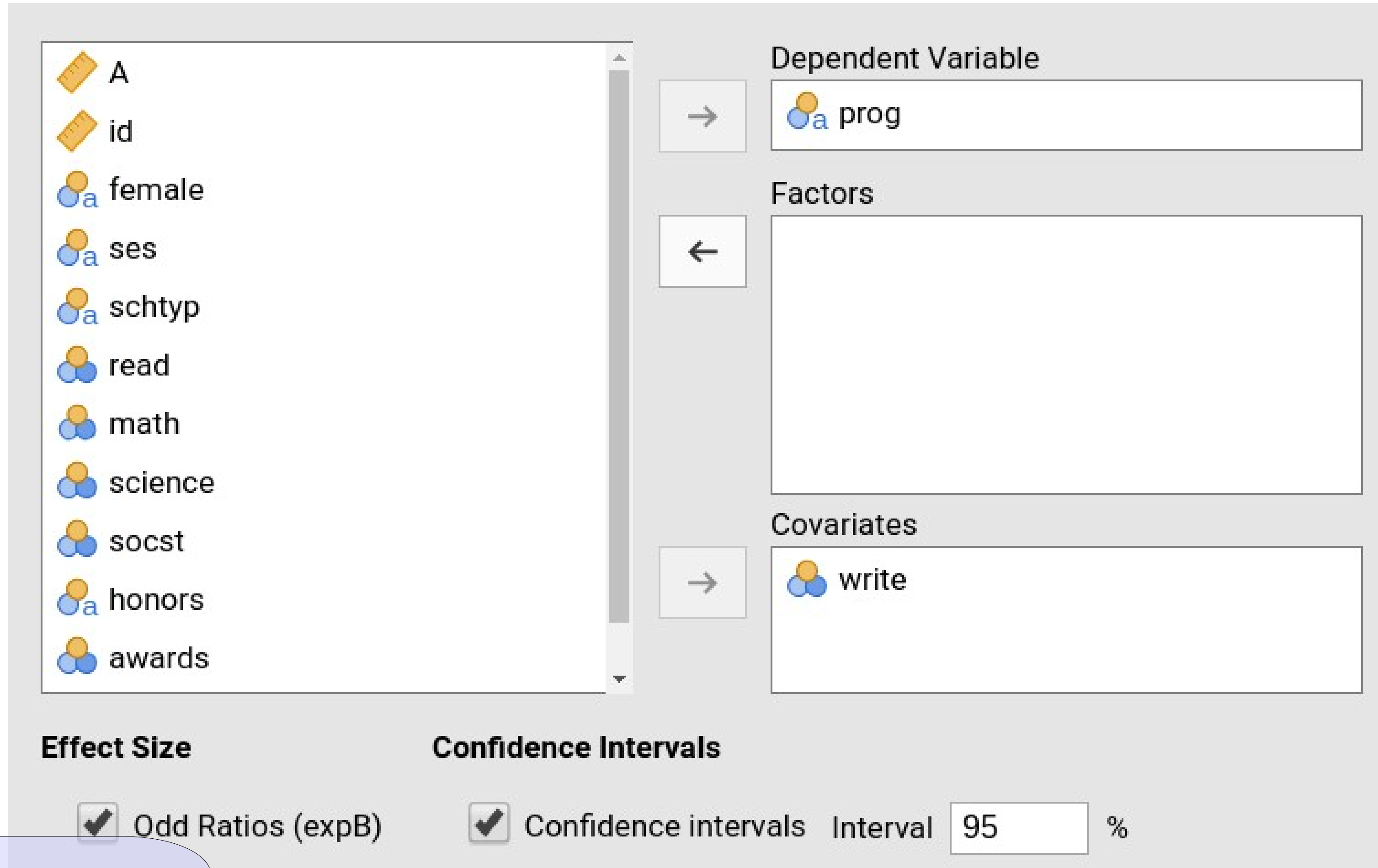
- the effect of the independent variables on the (log) odd of choosing **academic program over choosing general**,
- the (log) odd of choosing **vocation program over choosing general**

# Example




**Specify the type of model**

# Example




**Specify variables role**

# Results: model info

Model Info

| Info | Value | Comment |
|---|---|---|
| Model Type | Multinomial | Model for categorical y |
| Call | multinom | prog ~ 1 + write |
| Link function | logit | Log of the odd of each category over y=0 |
| Direction | P(y=x)/P(x=0) | P(prog=academic)/P(prog=general) , P(prog=vocation)/P(prog=general) |
| Distribution | Multinomial | Multi-event distribution of y |
| R-squared | 0.0911 | Proportion of reduction of error |
| AIC | 379.0217 | Less is better |
| Deviance | 371.0217 | Less is better |
| Residual DF | 4.0000 | |
| Converged | yes | A solution was found |

**Indicates the directions of the effects**

**The R-squared can be interpreted as usual**

# Results: estimates and tests

## Model Results

Indicates that there is an overall effect of the IV on the probabilities of being in the DV groups

The IV inluences the program choice

Loglikelihood ratio tests

| | $X^2$ | df | p |
|---|---|---|---|
| write | 37.2 | 2 | < .001 |

Fixed Effects Parameter Estimates

| | | | | | 95% Confidence Interval | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Response Contrasts | Names | Effect | Estimate | SE | Lower | Upper | exp(B) | z | p |
| academic - general | (Intercept) | (Intercept) | 0.7707 | 0.1849 | 0.4083 | 1.13312 | 2.161 | 4.168 | < .001 |
| | write | write | 0.0660 | 0.0210 | 0.0248 | 0.10717 | 1.068 | 3.143 | 0.002 |
| vocation - general | (Intercept) | (Intercept) | −0.0876 | 0.2265 | −0.5314 | 0.35627 | 0.916 | −0.387 | 0.699 |
| | write | write | −0.0518 | 0.0225 | −0.0959 | −0.00768 | 0.950 | −2.301 | 0.021 |

# Results: estimates and tests

## Model Results

Loglikelihood ratio tests

| | $X^2$ | df | p |
|---|---|---|---|
| write | 37.2 | 2 | < .001 |

Fixed Effects Parameter Estimates

| | | | | | 95% Confidence Interval | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Response Contrasts | Names | Effect | Estimate | SE | Lower | Upper | exp(B) | z | p |
| academic - general | (Intercept) | (Intercept) | 0.7707 | 0.1849 | 0.4083 | 1.13312 | 2.161 | 4.168 | < .001 |
| | write | write | 0.0660 | 0.0210 | 0.0248 | 0.10717 | 1.068 | 3.143 | 0.002 |
| vocation - general | (Intercept) | (Intercept) | −0.0876 | 0.2265 | −0.5314 | 0.35627 | 0.916 | −0.387 | 0.699 |
| | write | write | −0.0518 | 0.0225 | −0.0959 | −0.00768 | 0.950 | −2.301 | 0.021 |

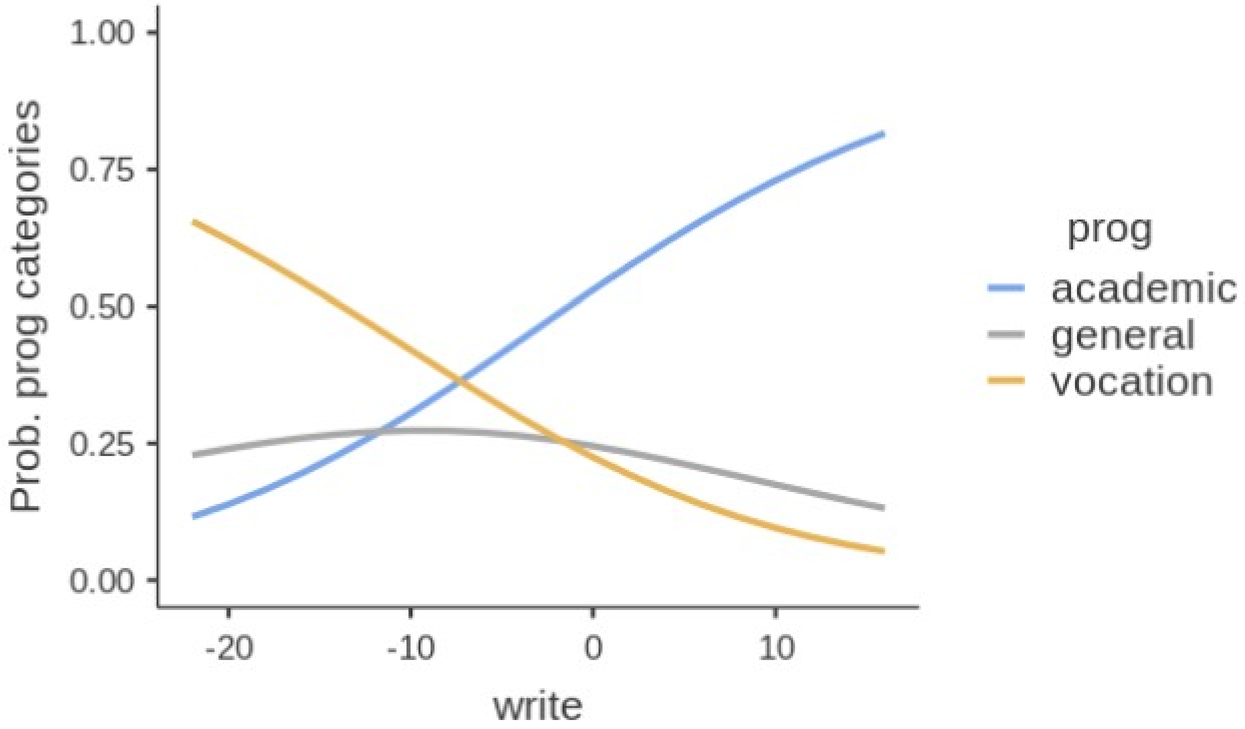
**Higher scores in write are associated with higher probability of going to academic rather than general**

**Higher scores in write are associated with lower probability of going to vocation rather than general**

# Results: plot

**The probability of being in each group defined by the DV**

**Plots**



# Generalized linear model

- Applying this logic we obtain a large set of possible statistical techniques

$$f(y_i) = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + .. b_k \cdot x_{ki} + e_i$$

| Dependent Variable | function | Distribution |
|---|---|---|
| Continuous | identity | Normal |
| Dichotomous | Logit of odd | Binomial |
| Categorical | Logit of odd | Multinomial |
| Ordinal | Cumulative Logit | Multinomial |
| Frequencies | Frequencies LN | Poisson |

Poisson model

Dependent variable is a count variable

# Count variable

A sample of children is measured with a test of aggression based on the school teachers evaluations. To understand the validity of the measure, a session of observed play is assessed, counting for each child how many aggressi behaviors s/he produces. The variable is the the count (the frequency) of aggressive acts produced by each child

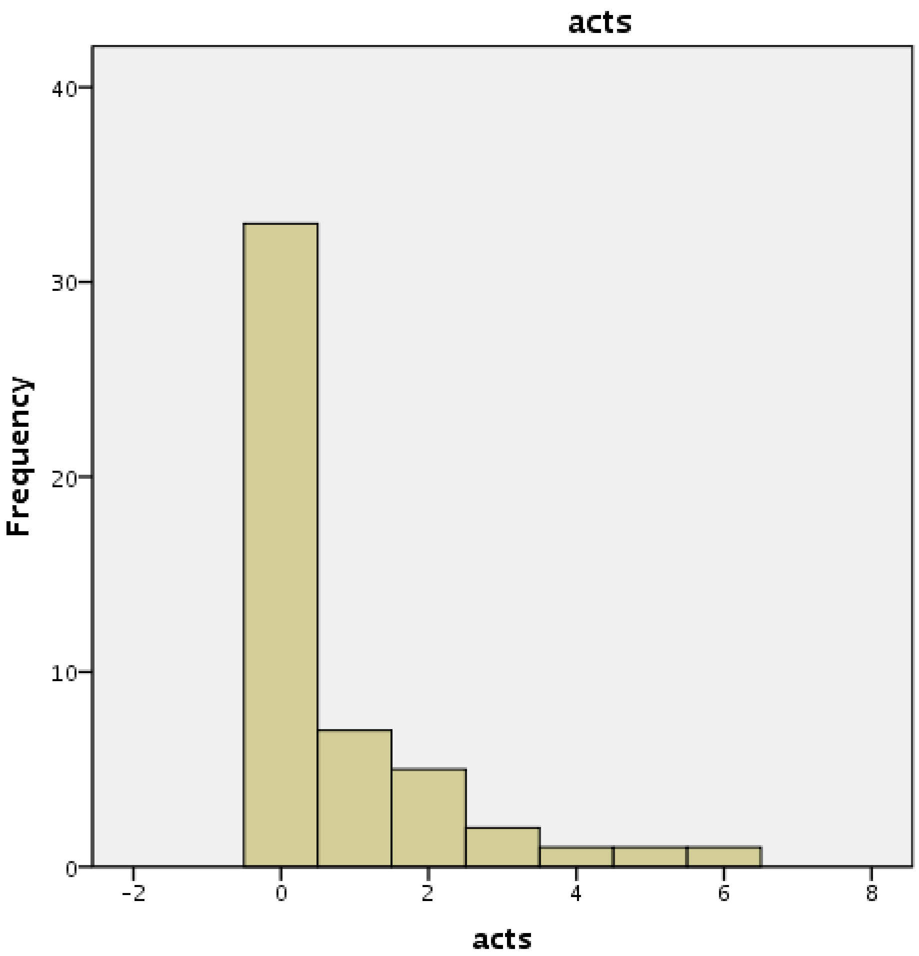
**acts**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 0 | 33 | 66.0 | 66.0 | 66.0 |
| | 1 | 7 | 14.0 | 14.0 | 80.0 |
| | 2 | 5 | 10.0 | 10.0 | 90.0 |
| | 3 | 2 | 4.0 | 4.0 | 94.0 |
| | 4 | 1 | 2.0 | 2.0 | 96.0 |
| | 5 | 1 | 2.0 | 2.0 | 98.0 |
| | 6 | 1 | 2.0 | 2.0 | 100.0 |
| | Total | 50 | 100.0 | 100.0 | |

# Count variable

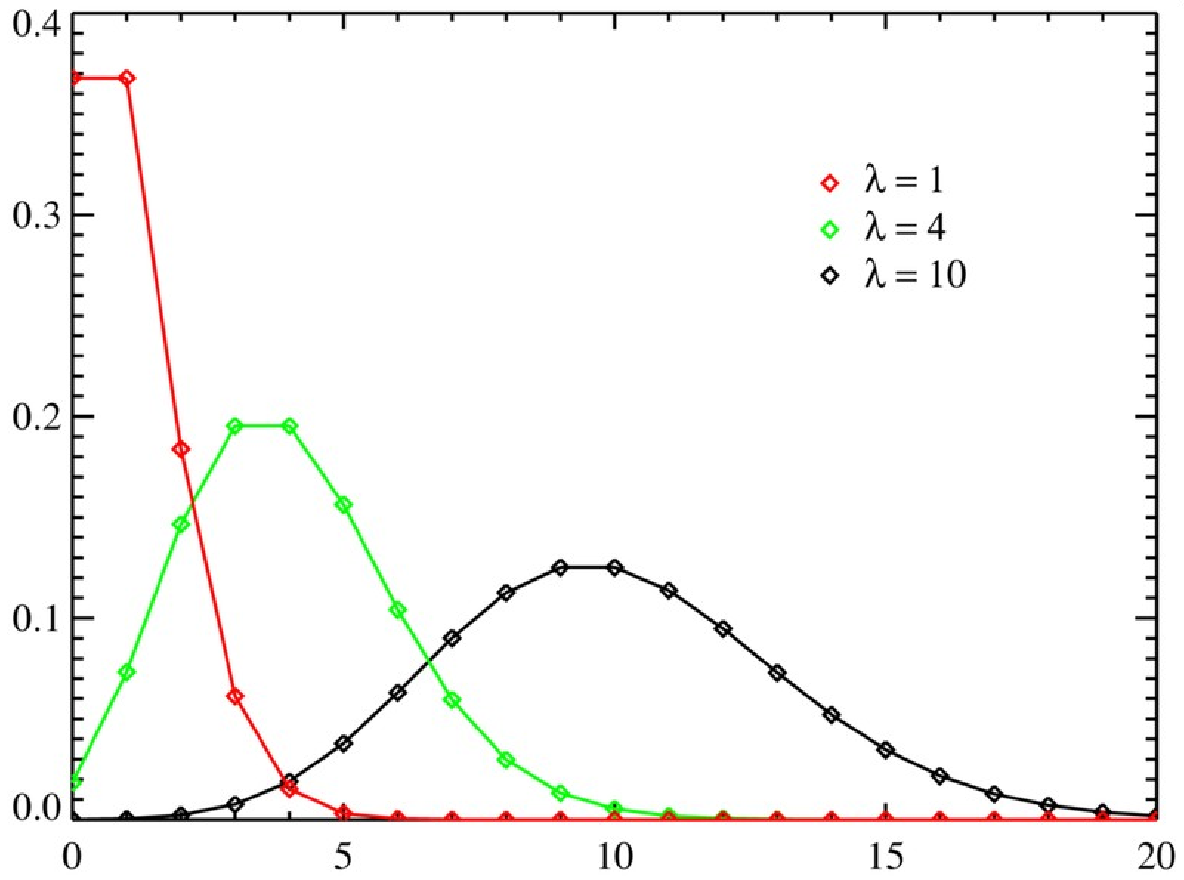
The dependent variable is clearly not normal.

Count data are likely to follow a **Poisson distribution**



# The Poisson distribution

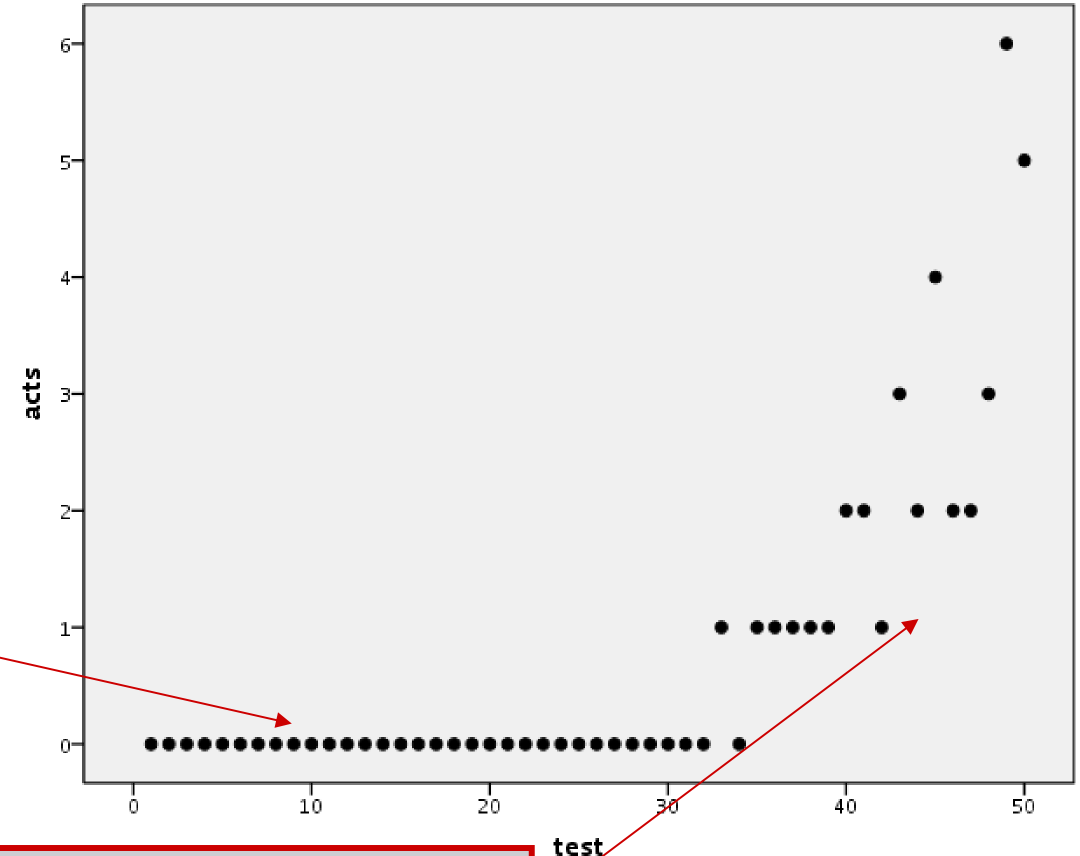
- The Poisson distribution describe the probability of observing an event with different possible frequency, given that the event has a fixed rate of occurring (λ)




The more the event is rare, the less the distribution resambles a normal distribition

# Relationships with counts variables

- Count variables distributed as Poisson tend to be related with other variables in a **non-linear** fashion



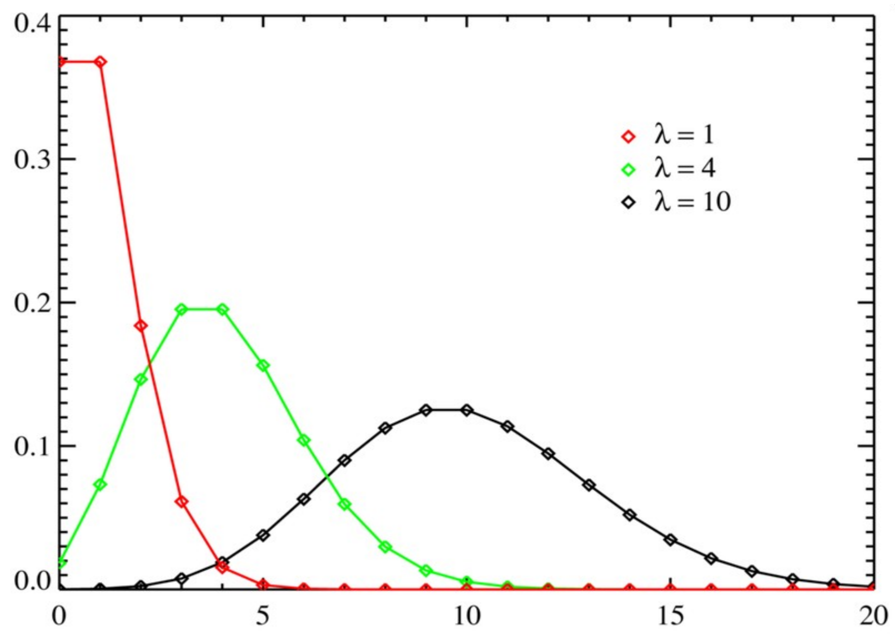

Low frequency in the count for a large range of the IV

Fast increase for a few high values of the IV

# Building a model

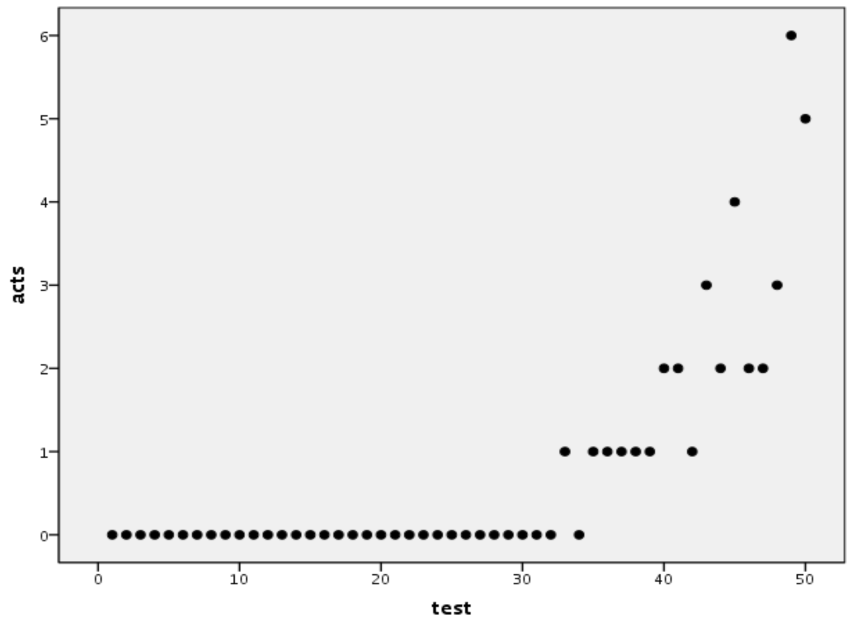
- Thus, we can be a GzLM assuming that the dependent variable follow a Poisson distribution

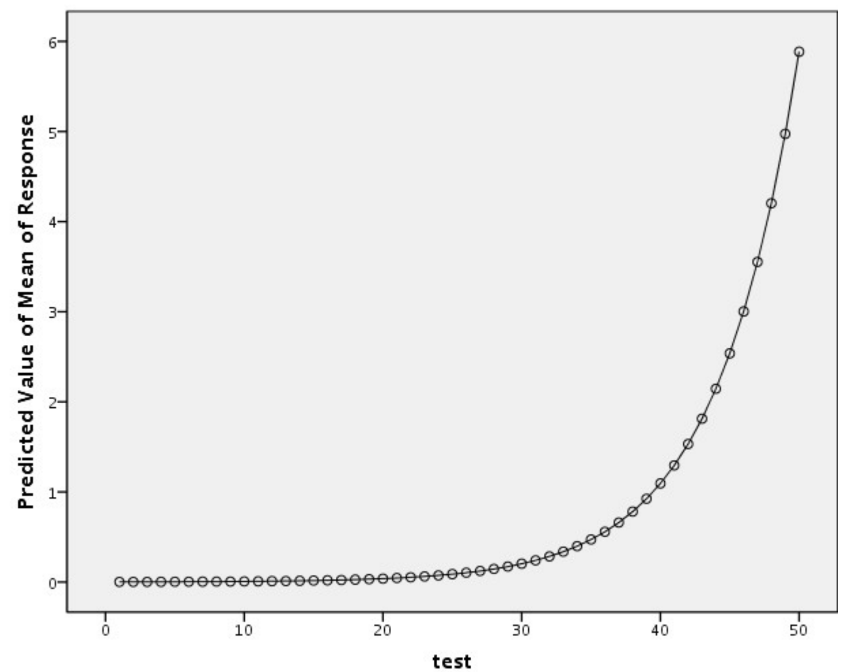
$$y = Poisson(y)$$



# Trasformation of “counts”

- To capture the non-linear shape of the relationship between the dependent variable and the indepenent variable(s) we use the logarithm transformation (link function)




$$y' = \ln(y)$$

# The Poisson model

- We end up with a GzLM with logarithm link function and Poisson distribution of error

Link function: logarithm

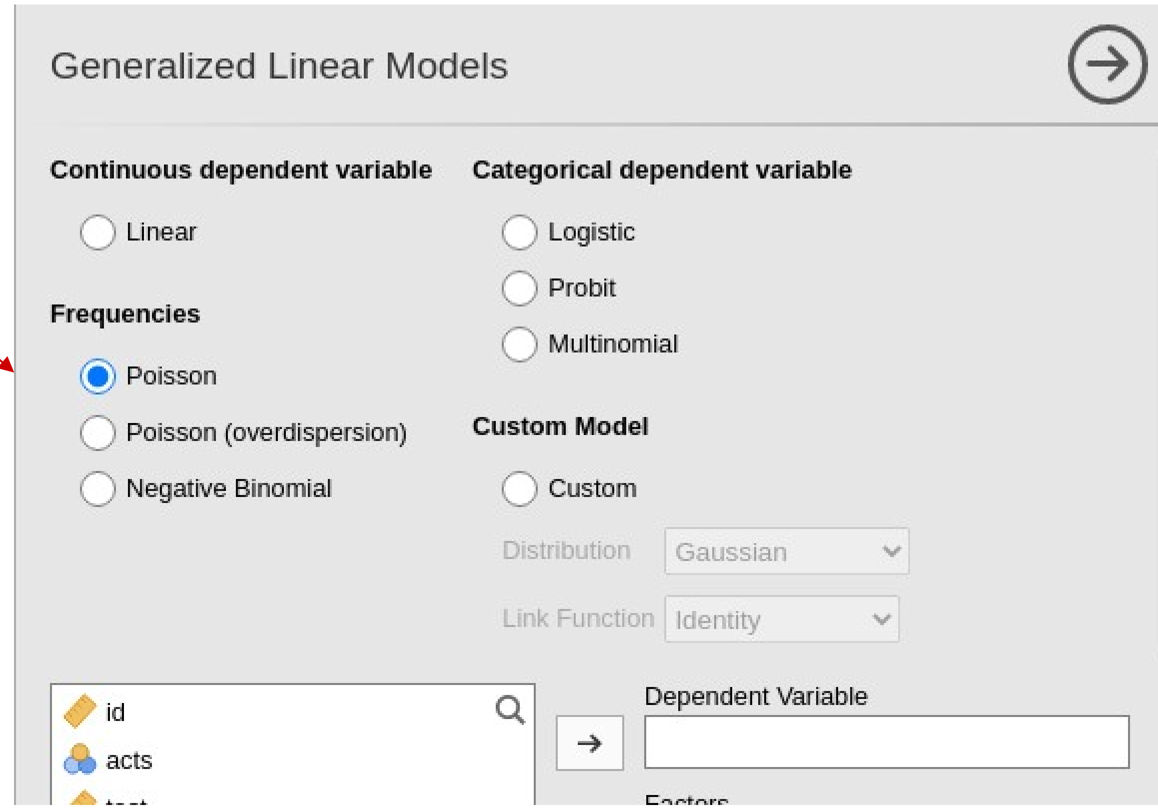
Error distribution: Poisson

$$\ln(y) = a + b_x x_i + b_w w_i + e_i$$

# The Poisson model: practice

- In jamovi we use the “generalized linear model” interface

Select Poisson loglinear

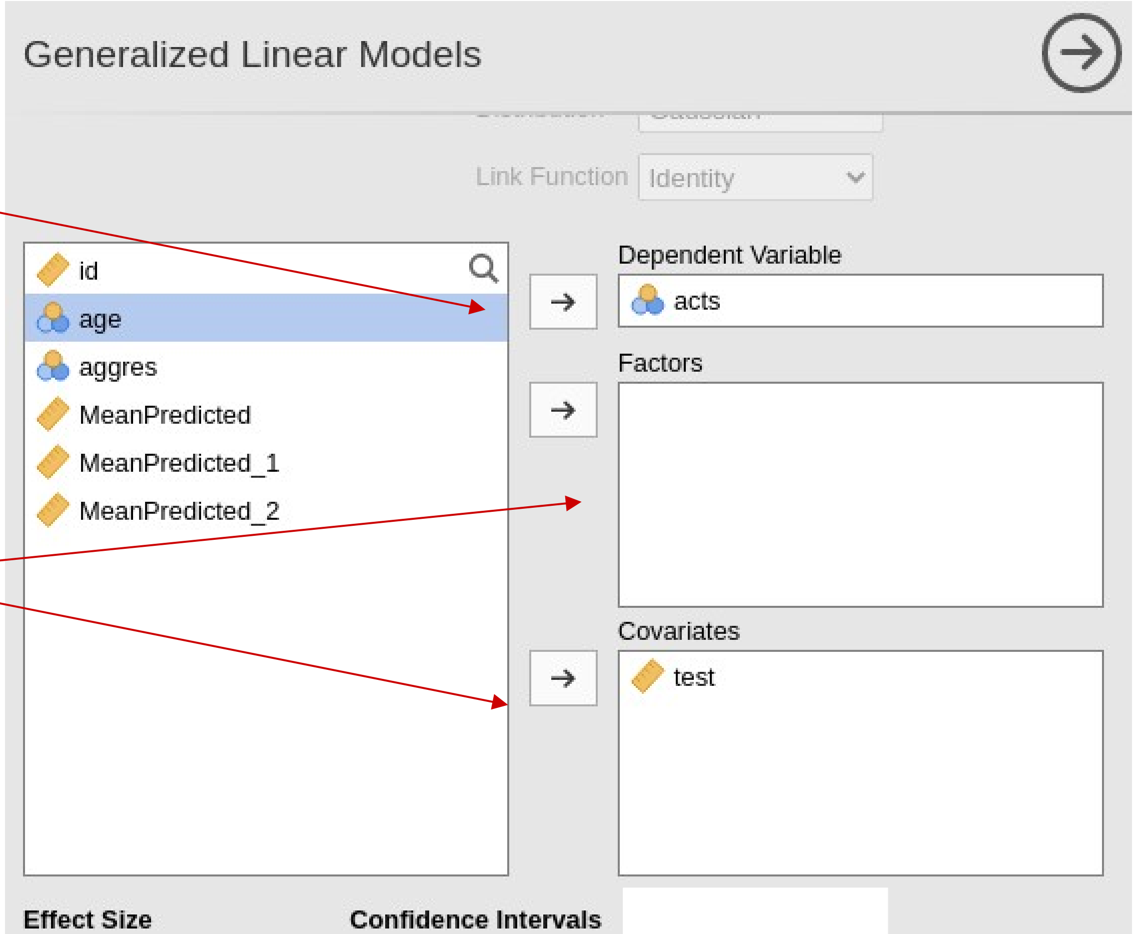


# The Poisson model: practice

- And follow the same steps we used for the logistic regression

Select the dependent variable

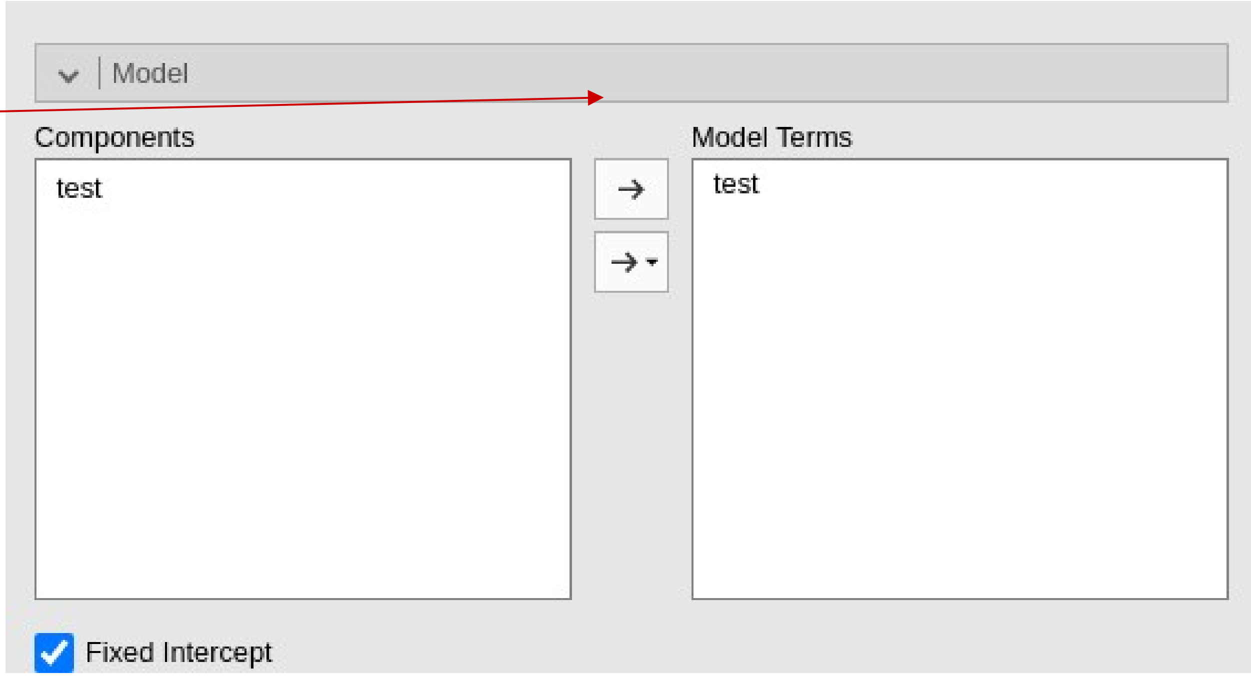
Define predictors



# The Poisson model: practice

- And follow the same steps we used for the logistic regression

Define effects




- For most models, the effects are set up automatically

# The Poisson model: practice

Model Info

| Info | Value | Comment |
|---|---|---|
| Model Type | Poisson | Model for count data |
| Call | glm | acts ~ 1 + test |
| Link function | log | Coefficients are in the log(y) scale |
| Distribution | Poisson | Model for count data |
| R-squared | 0.889 | Proportion of reduction of error |
| AIC | 58.129 | Less is better |

This tests the whole model
- (Like the F of the $R^2$ in GLM)

## Model Results

These test the effects
(Like the F of the effects in GLM)

Loglikelihood ratio tests

| | $X^2$ | df | p |
|---|---|---|---|
| test | 85.9 | 1 | < .001 |

Parameter Estimates

| | | | | 95% Exp(B) Confidence Interval | | | |
|---|---|---|---|---|---|---|---|
| Names | Estimate | SE | exp(B) | Lower | Upper | z | p |
| (Intercept) | −2.349 | 0.5492 | 0.0955 | 0.0280 | 0.245 | −4.28 | < .001 |
| test | 0.168 | 0.0275 | 1.1832 | 1.1269 | 1.256 | 6.11 | < .001 |

These are the coefficients

# Interpretation

Parameter Estimates

| Names | Estimate | SE | exp(B) | 95% Exp(B) Confidence Interval | | z | p |
|---|---|---|---|---|---|---|---|
| | | | | Lower | Upper | | |
| (Intercept) | −2.349 | 0.5492 | 0.0955 | 0.0280 | 0.245 | −4.28 | < .001 |
| test | 0.168 | 0.0275 | 1.1832 | 1.1269 | 1.256 | 6.11 | < .001 |

Logarithm scale: the increase of the logarithm of the frequency of DV for each unit more of the IV

# Exp(B)

- We can interpret the exp(B) which removes the log from the scale of B

Parameter Estimates

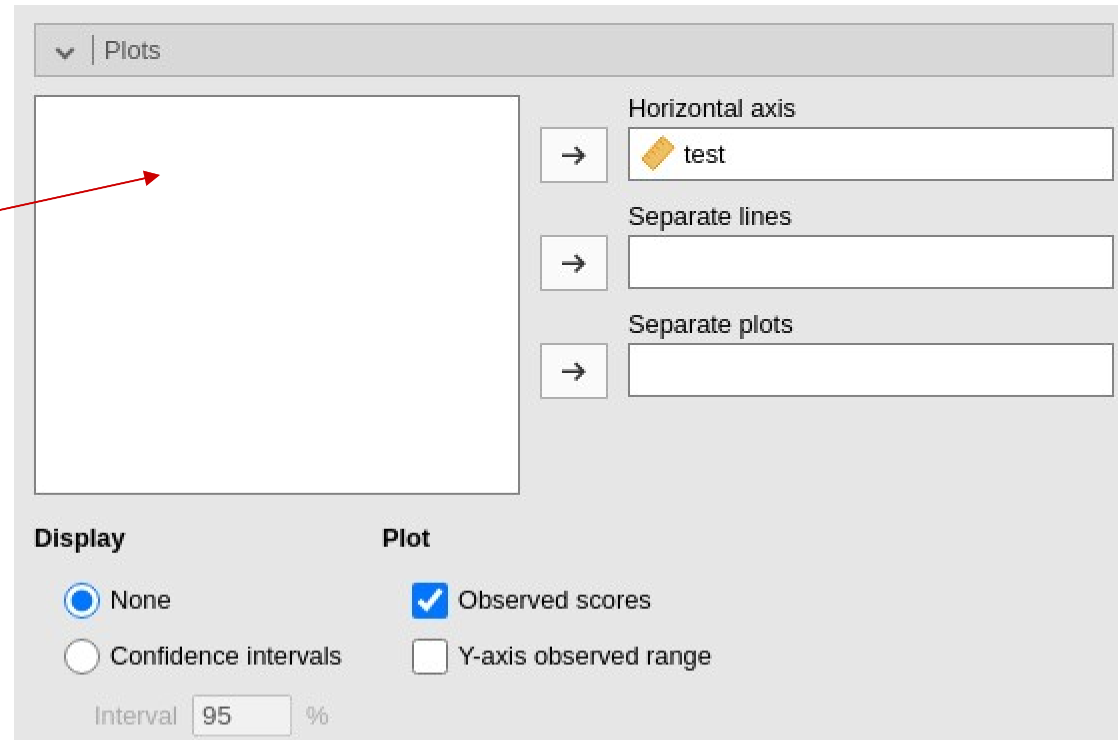
| Names | Estimate | SE | exp(B) | 95% Exp(B) Confidence Interval Lower | Upper | z | p |
|---|---|---|---|---|---|---|---|
| (Intercept) | −2.349 | 0.5492 | 0.0955 | 0.0280 | 0.245 | −4.28 | < .001 |
| test | 0.168 | 0.0275 | 1.1832 | 1.1269 | 1.256 | 6.11 | < .001 |

count scale: the rate of increase of the frequency of DV for each unit more of the IV

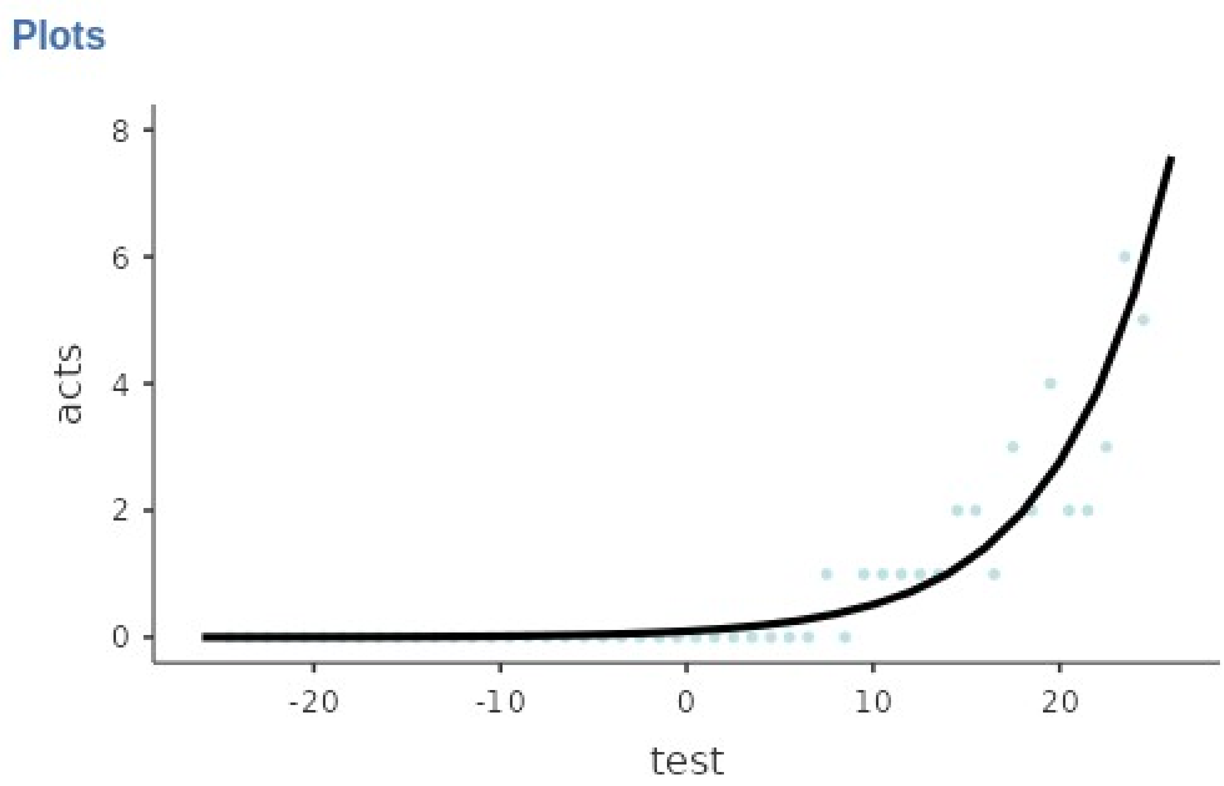
count scale: **how many times** the frequency of DV increases for each unit more of the IV

# Picture the model

Ask for the plot of the predicted values of the model



# And plot the predicted values



# Generalized Linear Mixed Models

# GzLM

**Statistical models aimed at studying the effects of one or more IV on dependent variables that are non-normally distributed**

**The mixed model generalize these models by allowing coefficients to vary across clusters of data**

# Theory

- The theory behind GGMM is simple: they are equivalent to the GGLM we just overviewed but the coefficients (intercepts and the effects) can be fixed and random, varying across clusters

- By specifying a model like we did in GGLM (for instance a logistic regression) a using the knowledge of the Mixed models (what can be random and what cannot be) we can specify and interpret a generalized mixed model.

# Practice

- From a practical point of view we use the SPSS interface of generalized mixed model which is, unfortunately, a bit strange!

# Example

- An experimental study on the relationship between mather and child. Children of this sample suffered from developmental difficulties. The sample features three categories of mothers: anxious, depressed, and control (no psychological condition).
- Mothers had to write an essay about the feelings and emotions they felt related to their child difficulties.
- Two essays were required, one about the feelings they felt thinking about the child past years, and one regarding the feelings they felt thinking about the future of the child.
- Essays were categorized by an independent coder as hostile or not hostile

# Example

- Research design is 3 GROUP (anxious, depressed, control) X 2 TIME (past vs future) , with the last factor as a repeated measure factor.
- The was also a measure of  Mental Development Index for the child, to be used as a covariate

# Data

- Data are in the long-format

| | ID | GROUP | Time | MDI | Hostility | var |
|---|---|---|---|---|---|---|
| 1 | 2010 | 1 | 0 | 87 | 1 | |
| 2 | 2010 | 1 | 1 | 87 | 1 | |
| 3 | 2023 | 1 | 0 | 78 | 1 | |
| 4 | 2023 | 1 | 1 | 78 | 1 | |
| 5 | 2029 | 1 | 0 | 84 | 1 | |
| 6 | 2029 | 1 | 1 | 84 | 1 | |
| 7 | 2130 | 1 | 0 | 97 | 0 | |
| 8 | 2130 | 1 | 1 | 97 | 0 | |
| 9 | 2131 | 2 | 0 | 72 | 0 | |
| 10 | 2131 | 2 | 1 | 72 | 1 | |
| 11 | 2291 | 2 | 0 | 99 | 1 | |
| 12 | 2291 | 2 | 1 | 99 | 0 | |
| 13 | 2344 | 2 | 0 | 99 | 0 | |
| 14 | 2344 | 2 | 1 | 99 | 1 | |
| 15 | 2345 | 1 | 0 | 95 | 0 | |
| 16 | 2345 | 1 | 1 | 95 | 1 | |
| 17 | 2426 | 1 | 0 | 118 | 1 | |
| 18 | 2426 | 1 | 1 | 118 | 1 | |
| 19 | 2601 | 1 | 0 | 92 | 1 | |
| 20 | 2601 | 1 | 1 | 92 | 1 | |
| 21 | 2666 | 2 | 0 | 106 | 0 | |
| 22 | 2666 | 2 | 1 | 106 | 1 | |
| 23 | 2691 | 1 | 0 | 102 | 0 | |

# Data

- Cross-tabs of the interesting variables

**Group * Time Crosstabulation**

Count

| | | Time | | Total |
|---|---|---|---|---|
| | | Past | Future | |
| Group | Control | 40 | 40 | 80 |
| | Anxiety | 48 | 48 | 96 |
| | Depression | 32 | 32 | 64 |
| Total | | 120 | 120 | 240 |

# Model

We define a logistic regression model with intercept as random to capture the dependency of the responses across participants

$$\ln\left(\frac{p}{1-p}\right) = \bar{a} + a_j + \bar{b}_1 \cdot Time_{ij} + \bar{b}_2 Group + \bar{b}_3 Time \cdot Group + e_{ij}$$

- Fixed effects? Intercept, group, time, and interaction effects
- Random effects? Intercepts
- Clusters? mothers

Exercise done in class

# jamovi GAMLj

- Imagine a study conducted in 70 schools. In each school the same exam is taken by students of equivalent age and grade. For each student, we recorded whether the student passed the exam, pass, the student's score in math test, math, and the number of extracurricular activities the student undertook during the semester.

- The researcher wants to estimate the effect of the math test on the probability of passing the exam, and also test whether the amount of extracurricular activities may moderate the math effect.

- Each school has a different number of students, ranging from 51 to 100. Each student presents three values: the score in the math test, the number of activity undertaken and whether the exam was passed pass=1 or not, pass=0.

# Design

- Schools are the clusters

## Frequencies

Frequencies of pass

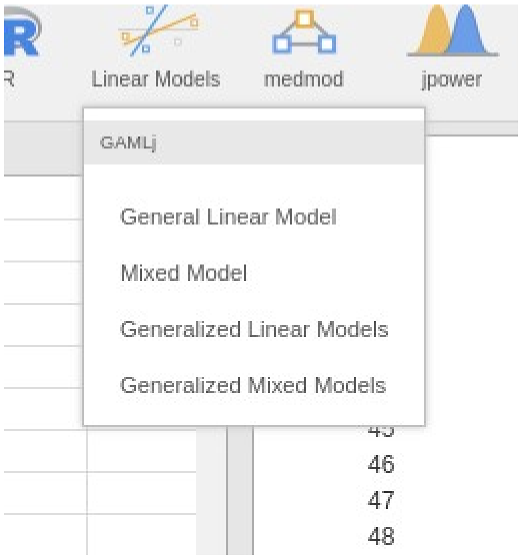
| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| 0 | 2479 | 49.2 % | 49.2 % |
| 1 | 2562 | 50.8 % | 100.0 % |

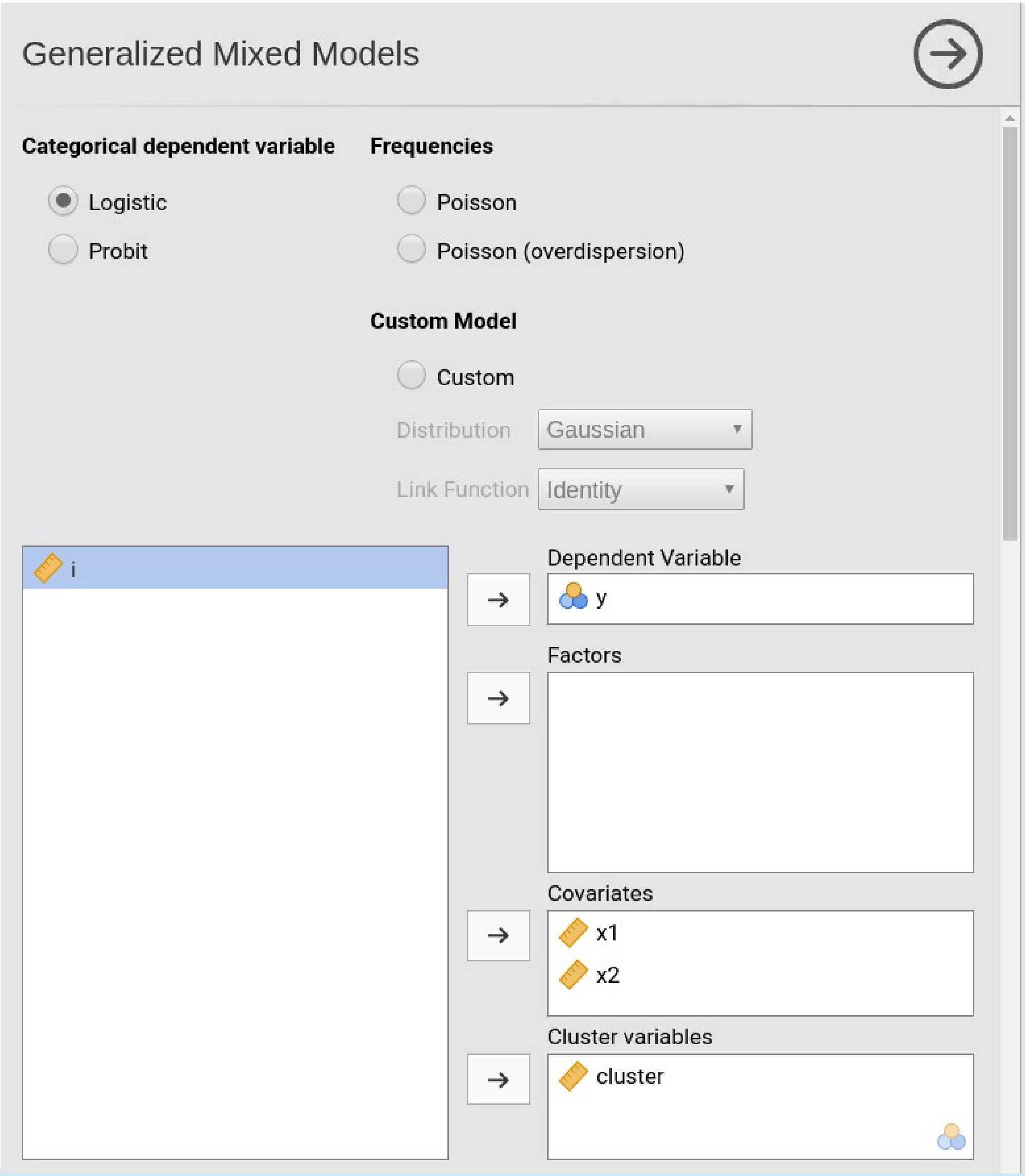
Frequencies of school

| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| 1 | 95 | 1.9 % | 1.9 % |
| 2 | 62 | 1.2 % | 3.1 % |
| 3 | 60 | 1.2 % | 4.3 % |
| 4 | 56 | 1.1 % | 5.4 % |
| 5 | 90 | 1.8 % | 7.2 % |
| 6 | 72 | 1.4 % | 8.6 % |
| 7 | 82 | 1.6 % | 10.3 % |
| 8 | 89 | 1.8 % | 12.0 % |
| 9 | 100 | 2.0 % | 14.0 % |
| 10 | 59 | 1.2 % | 15.2 % |

# GzLMM

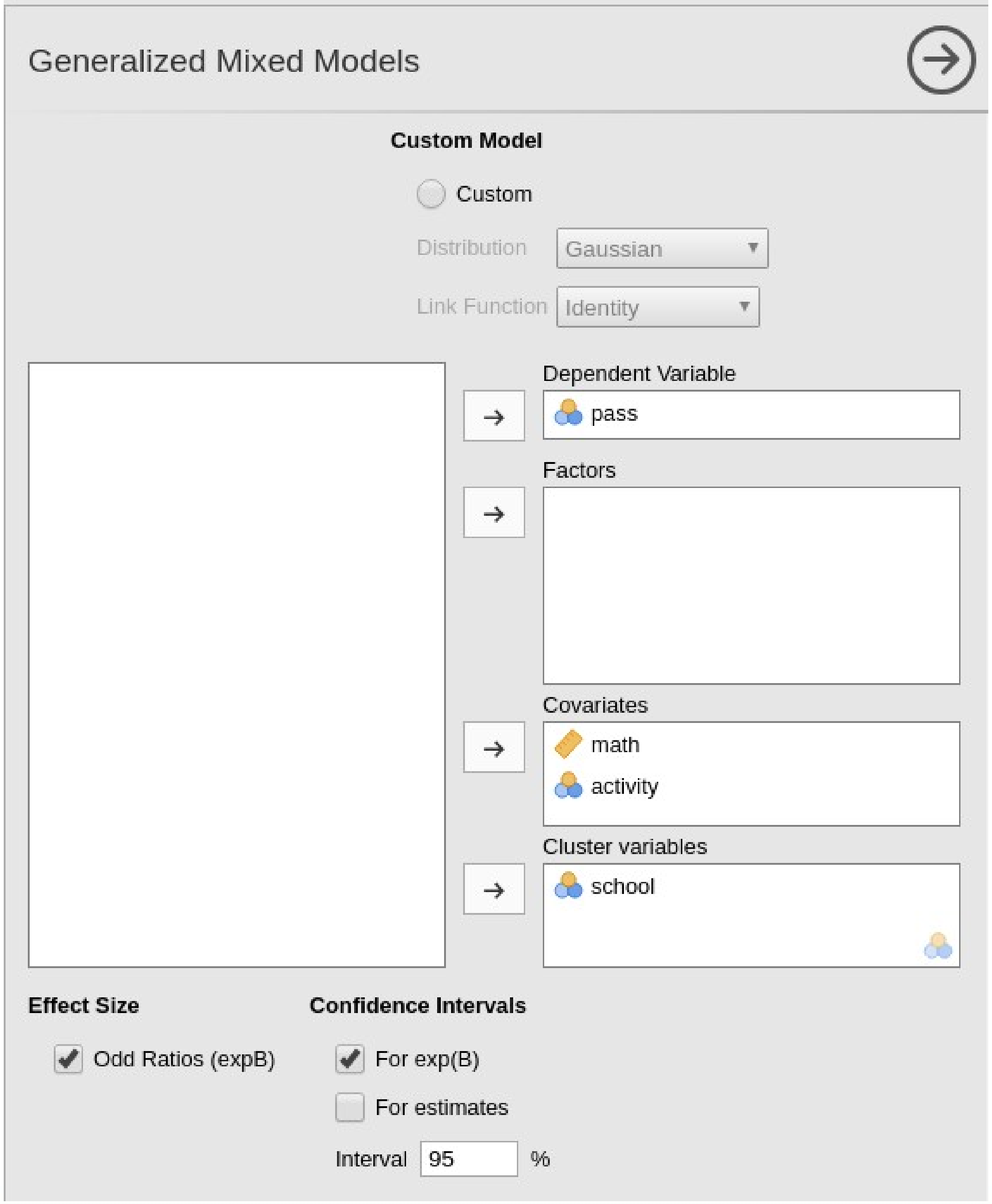
- We launch the module



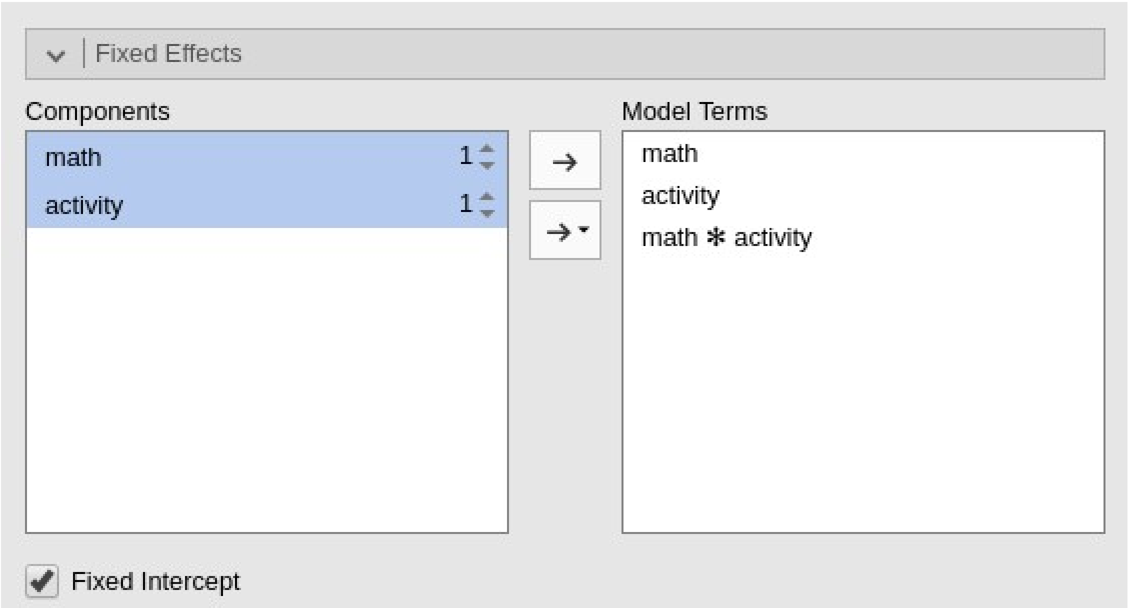
# GzLMM

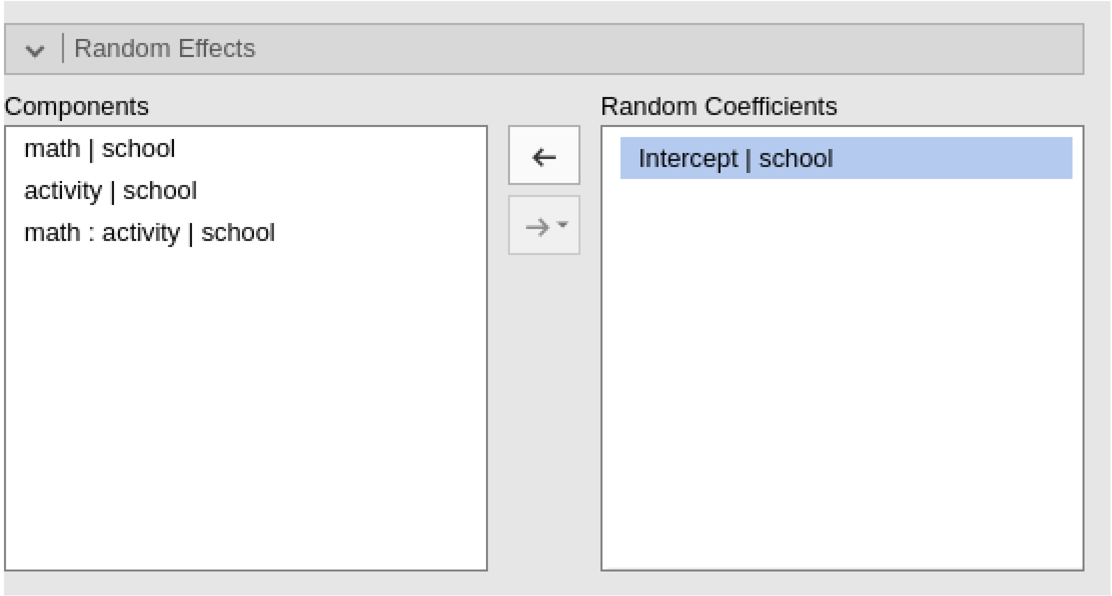
- We select the variables role



# GzLMM

- Define the model parameters



# Results

- Info table

Model Info

| Info | Value | Comment |
|---|---|---|
| Model Type | Logistic | Model for binary y |
| Call | glm | pass ~ 1 + math + activity + math:activity + (1 \| school) |
| Link function | Logit | Log of the odd of y=1 over y=0 |
| Direction | P(y=1)/P(y=0) | P( pass = 1 ) / P( pass = 0 ) |
| Distribution | Binomial | Dichotomous event distribution of y |
| LogLikel. | −2785.0640 | Less is better |
| R-squared | 0.0395 | Marginal |
| R-squared | 0.3787 | Conditional |
| AIC | 5580.1300 | Less is better |
| BIC | 5612.7547 | Less is better |
| Deviance | 5287.0900 | Conditional |
| Residual DF | 5036.0000 | |

[3]

**Direction of the model: What are we predicting?**

**R-squared for the whole model and for the fixed effects**

# Results

- Random component

Random Components

| Groups | Name | SD | Variance |
|---|---|---|---|
| school | (Intercept) | 1.34 | 1.80 |
| Residuals | | 1.00 | 1.00 |

*Note*. Number of Obs: 5041 , groups: school , 70

**Residual variance in always 1**

# Results

- Fixed effects

## Model Results

Fixed Effect Omnibus tests

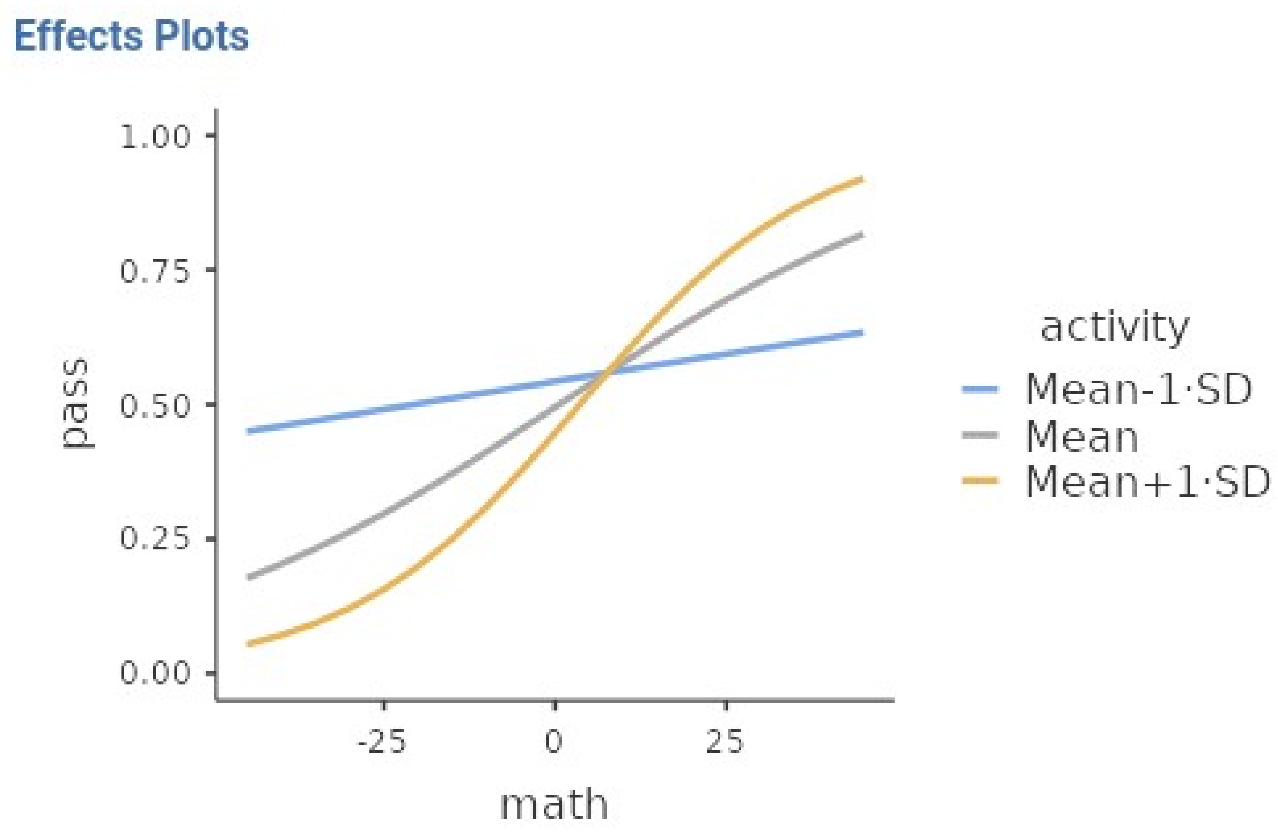
| | $X^2$ | df | p |
|---|---|---|---|
| math | 95.125 | 1.000 | < .001 |
| activity | 31.939 | 1.000 | < .001 |
| math ✻ activity | 52.336 | 1.000 | < .001 |

Fixed Effects Parameter Estimates

| | | | | 95% Exp(B) Confidence Interval | | | |
|---|---|---|---|---|---|---|---|
| Names | Estimate | SE | exp(B) | Lower | Upper | z | p |
| (Intercept) | −0.020 | 0.164 | 0.980 | 0.710 | 1.352 | −0.123 | 0.902 |
| math | 0.034 | 0.003 | 1.034 | 1.027 | 1.041 | 9.753 | < .001 |
| activity | −0.186 | 0.033 | 0.830 | 0.779 | 0.886 | −5.651 | < .001 |
| math ✻ activity | 0.024 | 0.003 | 1.025 | 1.018 | 1.032 | 7.234 | < .001 |

# Results

- Plots



# Results

- Fixed effects

## Model Results

Fixed Effect Omnibus tests

| | $X^2$ | df | p |
|---|---|---|---|
| math | 95.1 | 1.00 | < .001 |
| activity | 31.9 | 1.00 | < .001 |
| math ✻ activity | 52.3 | 1.00 | < .001 |

**GAMLj uses the Chi-Squared**

# Results

- Fixed effects

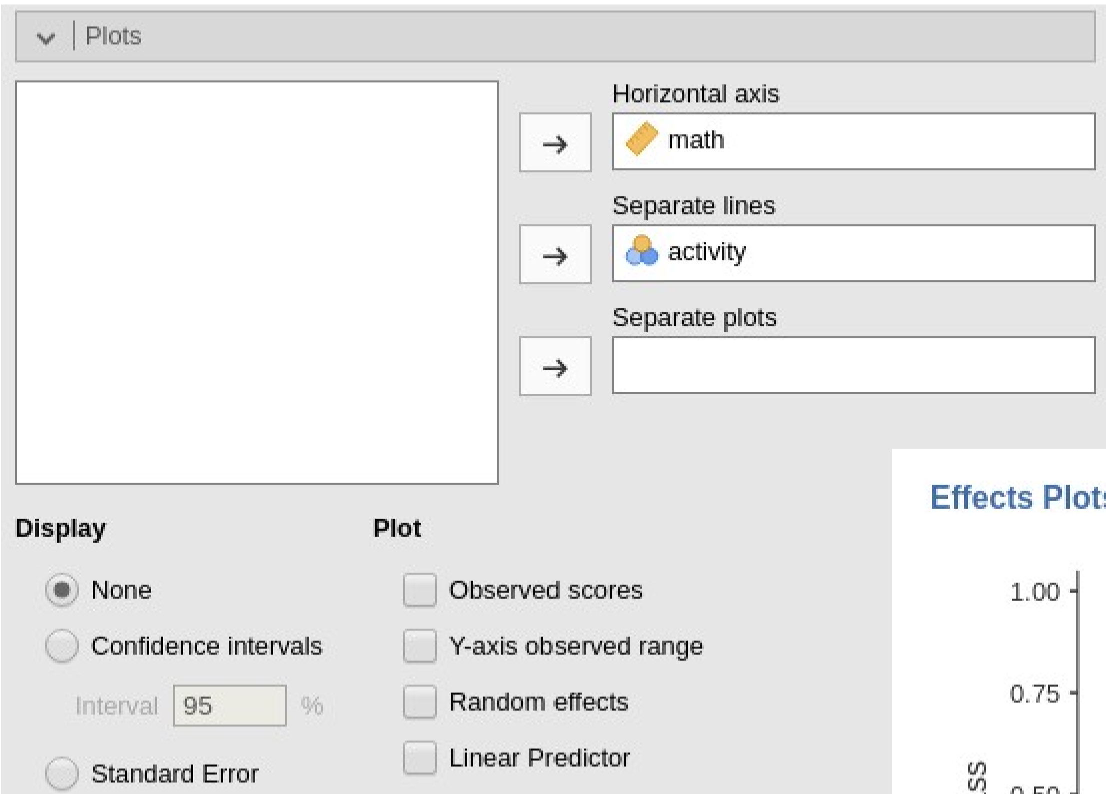
Here we found the exp(B)

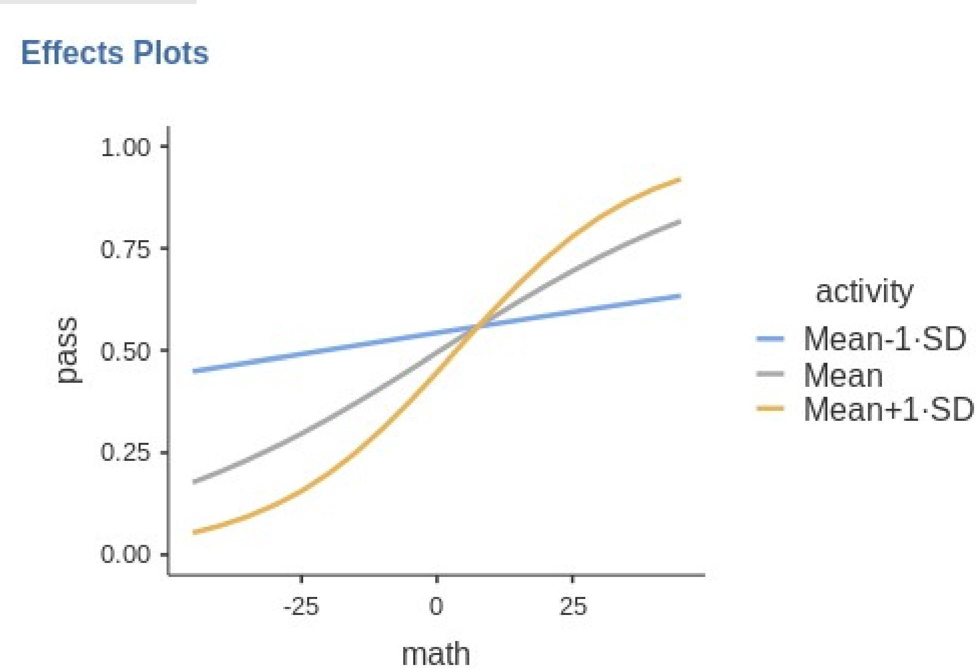
Fixed Effects Parameter Estimates

| Names | Estimate | SE | exp(B) | 95% Exp(B) Confidence Interval Lower | Upper | z | p |
|---|---|---|---|---|---|---|---|
| (Intercept) | −0.0202 | 0.16416 | 0.980 | 0.710 | 1.352 | −0.123 | 0.902 |
| math | 0.0337 | 0.00345 | 1.034 | 1.027 | 1.041 | 9.753 | < .001 |
| activity | −0.1858 | 0.03288 | 0.830 | 0.779 | 0.886 | −5.651 | < .001 |
| math ✻ activity | 0.0245 | 0.00338 | 1.025 | 1.018 | 1.032 | 7.234 | < .001 |

# Results

- Plot



# Recap

- **General linear model** allows for analyzing a variety of design with normally distributed DV by apply regression/ANOVA tecniques
- For repeated measures (or in general dependent data), we use the **Linear Mixed model** to allow coefficients to vary randomly across clusters, thus taking dependency into the account
- When the DV is categorical, we can use the **Generalized Linear model** which allows to apply regression/ANOVA tecniques to categorical dependent variables
- For repeated measures (or in general dependent data), we use the **Generalized Linear Mixed model** to allow coefficients to vary randomly across clusters, thus taking dependency into the account


# The End