Generalized Linear Models with `brms`

Filippo Gambarota

Recap about linear models

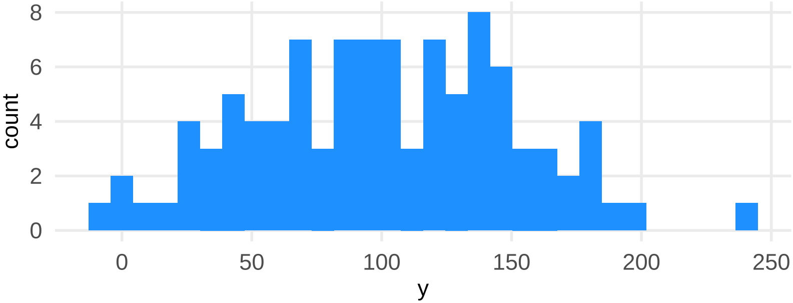
(almost) everything is a linear model

Most of the statistical analysis that you usually perfom, is essentially a linear model.

- The **t-test** is a linear model where a numerical variable `y` is predicted by a factor with two levels `x`
- The **one-way anova** is a linear model where a numerical variable `y` is predicted by one factor with more than two levels `x`
- The **correlation** is a linear model where a numerical variable `y` is predicted by another numerical variable `x`
- The **ancova** is a linear model where a numerical variable `y` is predicted by a numerical variable `x` and a factor with two levels `g`
- ...

What is a linear model?

Let's start with a single variable y. We assume that the variable comes from a Normal distribution:



What is a linear model?

What we can do with this variable? We can estimate the parameters that define the Normal distribution thus μ (the mean) and σ (the standard deviation).

```
mean(y)
#> [1] 100
sd(y)
#> [1] 50
```

What is a linear model?

Using a linear model we can just fit a model without predictors, also known as intercept-only model.

```r
fit <- glm(y ~ 1, family = gaussian(link = "identity"))
summary(fit)
```

```
#> 
#> Call:
#> glm(formula = y ~ 1, family = gaussian(link = "identity'
#> 
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)      100          5      20   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
#> 
#> (Dispersion parameter for gaussian family taken to be 25
#> 
#>     Null deviance: 247500  on 99  degrees of freedom
```

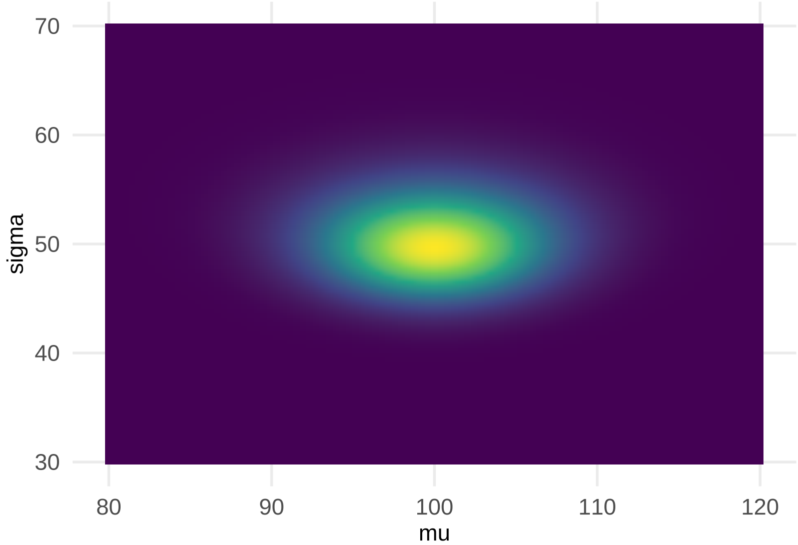
What is a linear model?

I am using `glm` because I want to estimate parameters using Maximul Likelihood, but the results are the same as using `lm`.

Basically we estimated the mean `(Intercept)` and the standard deviation `Dispersion`, just take the square root thus 50.

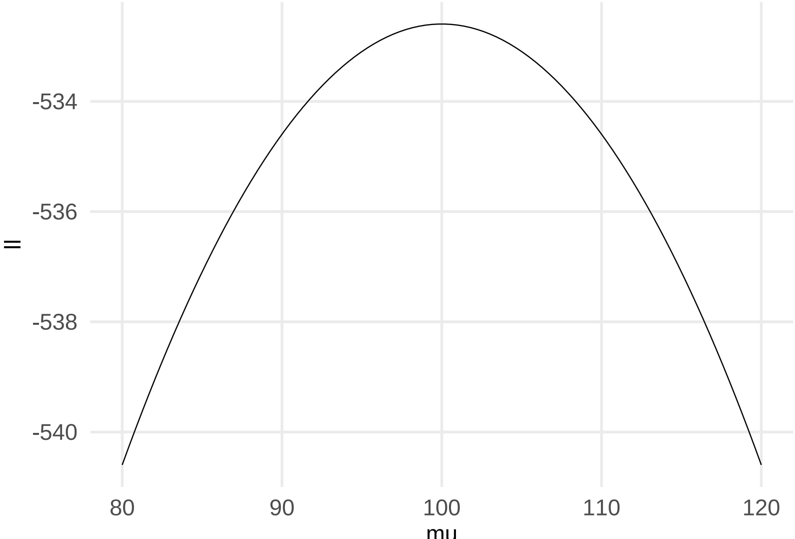
What we are doing is essentially finding the μ and σ that maximised the log-likelihood of the model fixing the observed data.

What is a linear model?



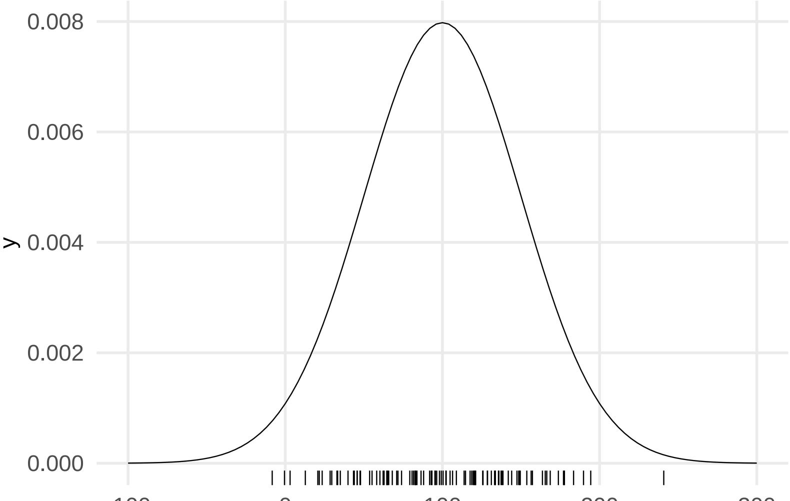
What is a linear model?

And assuming that we know σ (thus fixing it at 50):



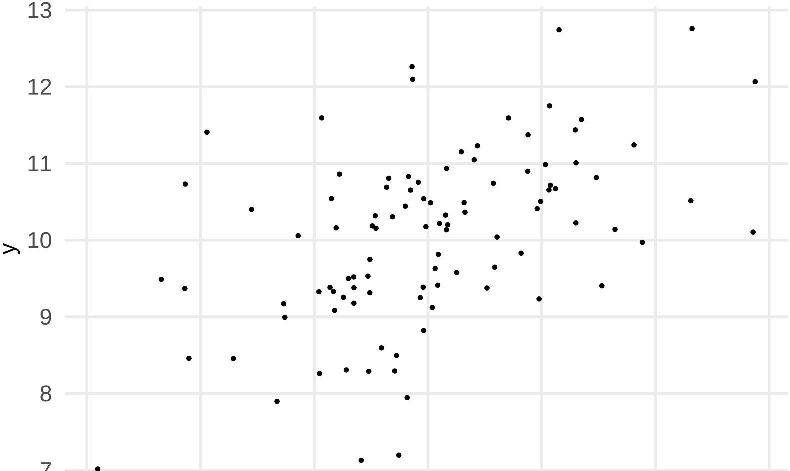
What is a linear model?

Thus, with the estimates of `glm`, we have this model fitted on the data:



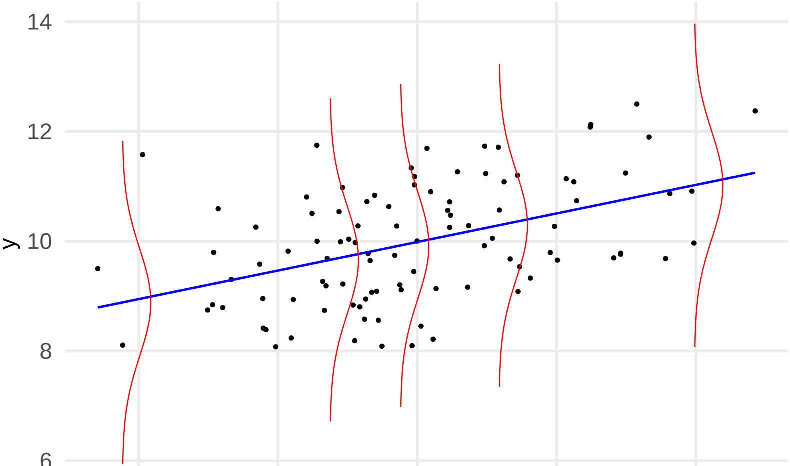
Including a predictor

When we include a predictor, we are actually try to explain the variability of `y` using a variable `x`. For example, this is an hypothetical relationship:



Assumptions of the linear model

More practicaly, we are saying that the model allows for varying the mean i.e., each `x` value can be associated with a different μ but with a fixed (and estimated) σ.



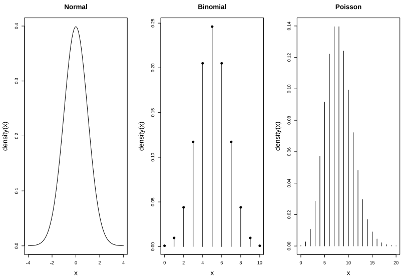
Generalized linear models

Recipe for a GLM

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

Random Component

The **random component** of a GLM identify the response variable y coming from a certain probability distribution.



Random Component

- In practice, by definition the GLM is a model where the random component is a distribution of the Exponential Family. For example the Gaussian distribution, the Gamma distribution or the Binomial are part of the Exponential Family.
- These distribution can be described using a **location** parameter (e.g., the mean) and a **scale** parameter (e.g., the variance).
- The distributions are defined by parameters (e.g., μ and σ for the Gaussian or λ for the Poisson). The location (or mean) can be directly one of the parameter or a combination of parameters.

Random Component, Poisson example

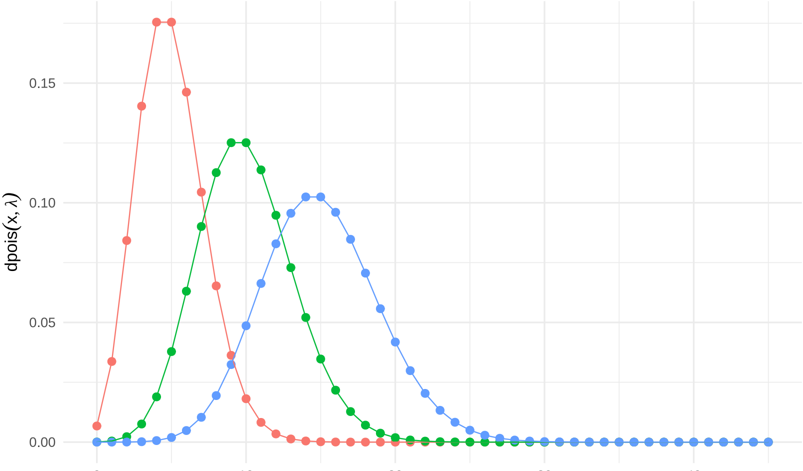
For example, the Poisson distribution is defined as:

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

Where k is the number of events and λ (the only parameter) is the *rate*.

Random Component, Poisson example

The mean or location of the Poisson is λ and also the scale or variance is λ. Compared to the Gaussian, there are no two parameters.



Random Component

To sum-up, the random component represents the assumption about the nature of our response variable. **With GLM we want to include predictors to explain *systematic* changes of the mean (but also the scale/variance) of the random component**.

Assuming a Gaussian distribution, we try to explain how the mean of the Gaussian distribution change according to our predictors. For the Poisson, we include predictors on the λ parameters for example.

The Random Component is called random, beacause it determines how the **error term** ϵ of our model is distributed.

Systematic Component

The systematic component of a GLM is the combination of predictors (i.e., independent variables) that we want to include in the model.

The systematic component is also called *linear predictor* η and is usually written in equation terms as:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Note that I am omitting the $+\epsilon_i$ that you usually find at the end because this is the combination of predictors without errors.

Systematic Component, an example

Assuming that we have two groups and we want to see if there are differences in a depression score. This is a t-test, or better a linear model, or better a GLM.

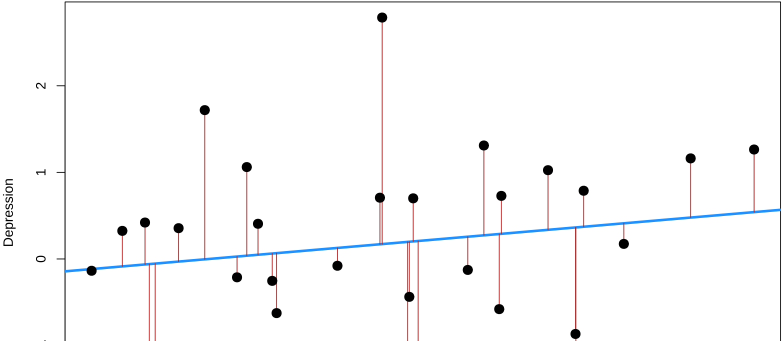
Ignoring the random component, we can have a systematic component written in this way:

$$\eta_i = \beta_0 + \beta_1 \mathsf{group}_i$$

Assuming that the group is dummy-coded, β_0 is the mean of the first group and β_1 is the difference between the two groups. In other terms, these are the true or estimated values without the error (i.e., the random component).

Systematic Component, an example

Another example, assuming we have the same depression score and we want to predict it with an anxiety score. The blue line is the true/estimated regression line where η_i is the expected value for the observation x_i. The red segments are the errors or residuals i.e., the random component.



Systematic Component

To sum-up, the systematic component is the combination of predictors that are used to predict the mean of the distribution that is used as random component. The errors part of the model is distributed as the random component.

Link Function

The final element is the **link function**. The idea is that we need a way to connect the systematic component η to the random component mean μ.

The **link function** $g(\mu)$ is an **invertible** function that connects the mean μ of the random component with the *linear combination* of predictors.

Thus $\eta_i = g(\mu_i)$ and $\mu_i = g(\eta_i)^{-1}$. The systematic component is not affected by $g()$ while the relationship between μ and η changes using different link functions.

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

$$\mu_i = g(\eta_i)^{-1} = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$