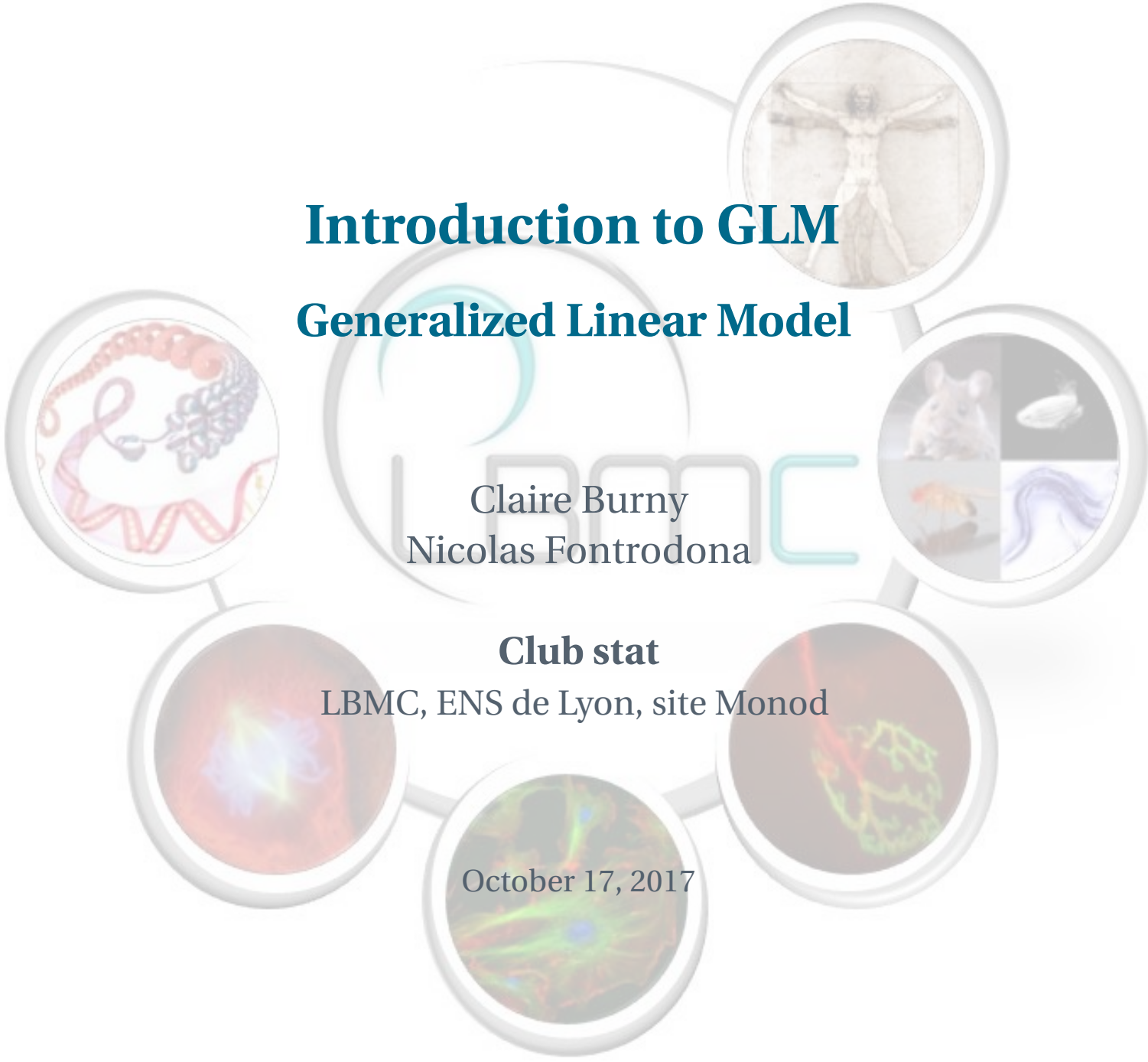


Introduction to GLM

Generalized Linear Model

Claire Burny
Nicolas Fontrodona

Club stat

LBMC, ENS de Lyon, site Monod

October 17, 2017

Definition

The **generalized linear models and linear models**, allow to study the relation between the response variable (Y) and a sets of explanatory variables ($X_1 ... X_k$)

The linear models are composed of:

- A response variable (Y) - Variable of interest
  - Let's say that ($Y_1 ... Y_n$) is a sample of size n of Y. $Y_1 ... Y_n$ are independant.
    - Y_i is normally distributed
- Explanatory variable(s) ($X_1 ... X_k$) - Variable(s) used to explain the variability in the response variable
- Explanatory variables can be expressed as : $\beta_0 + \beta_1 X_1 + ... + \beta_k X_k$
- Sometimes, an explanatory variable X_j can be deduced by elementary variables.
  - $X_3 = X_1 * X_2$

Linear models

More precisely, linear models can be expressed as :

$$E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon$$

Where :

- $E(Y)$ is the expected value of Y
- ϵ is the error parameter (must follow a normal distribution and homoscedastic)

We want to find the equation that best suits our data $(Y_1 ... Y_n)$. The parameters $\beta_0, \beta_1 ... \beta_n$ can be estimated by the least-square method. Their estimations are those which minimize:

$$\Sigma_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}))^2$$

Limits of linear models

- Can't explain a response variable that don't follow a normal distribution
- Can't explain a response variable that takes value in a particular interval
- Explanatory variables must have a linear effect on the response variable

To overcome those issues, we can use a generalized linear model

Distribution	*Interval*	*Uses*	*link function*
Normal	$]-\infty,+\infty[$	Linear response data	$E(Y)=\beta X$
Poisson	$[0,+\infty[$	Count data	$log(E(Y))=\beta X$
Bernoulli	$\{0,1\}$	outcome of an event	$log\left(\frac{E(Y)}{1-E(Y)}\right)=\beta X$
Binomial	$\{0,...,N\}$	outcome of N events	$log\left(\frac{E(Y)}{1-E(Y)}\right)=\beta X$
Exponential/Gamma	$]-\infty,+\infty[$	Exponential response data	$E(Y)^{-1}=\beta X$

Poisson distribution

Count data

Y is a categorial continuous data. Let's note $E(Y) = \mu$.

- $Y \hookrightarrow Pois(\mu)$ with $P(Y = k) = \frac{\mu^k e^{-\mu}}{k!}$
- $E(Y) = Var(Y) = \mu$

Poisson distribution

Count data

Y is a categorial continuous data. Let's note $E(Y) = \mu$.

- $Y \hookrightarrow Pois(\mu)$ with $P(Y = k) = \frac{\mu^k e^{-\mu}}{k!}$
- $E(Y) = Var(Y) = \mu$
- The link function is the log, the model is:

$$log(\mu) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \in]-\infty, +\infty[$$

Poisson distribution

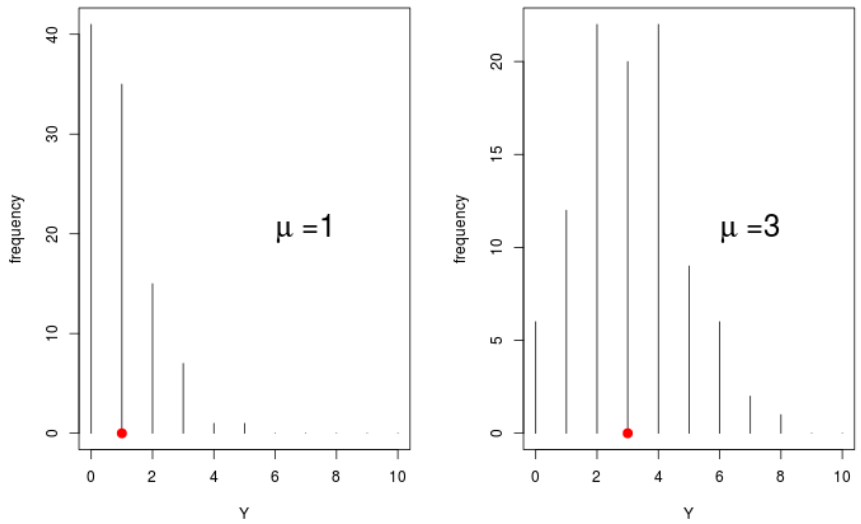
Count data

Y is a categorial continuous data. Let's note $E(Y) = \mu$.

- $Y \hookrightarrow Pois(\mu)$ with $P(Y = k) = \frac{\mu^k e^{-\mu}}{k!}$
- $E(Y) = Var(Y) = \mu$
- The link function is the log, the model is:

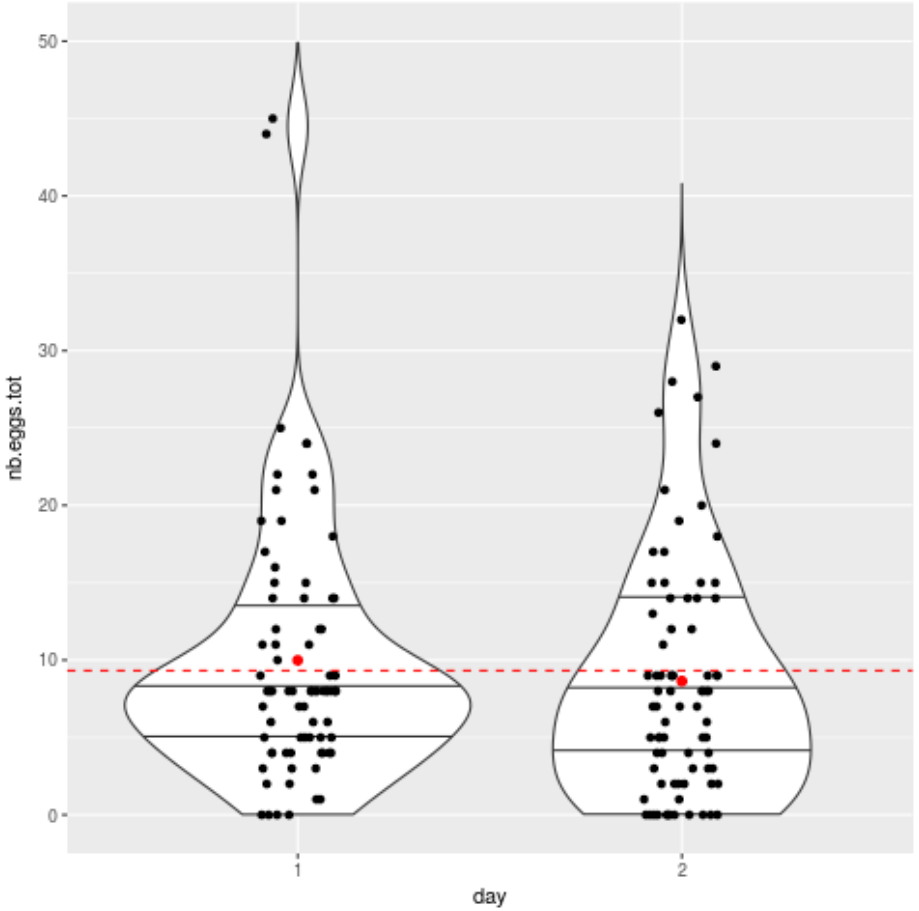
$$log(\mu) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \in]-\infty, +\infty[$$

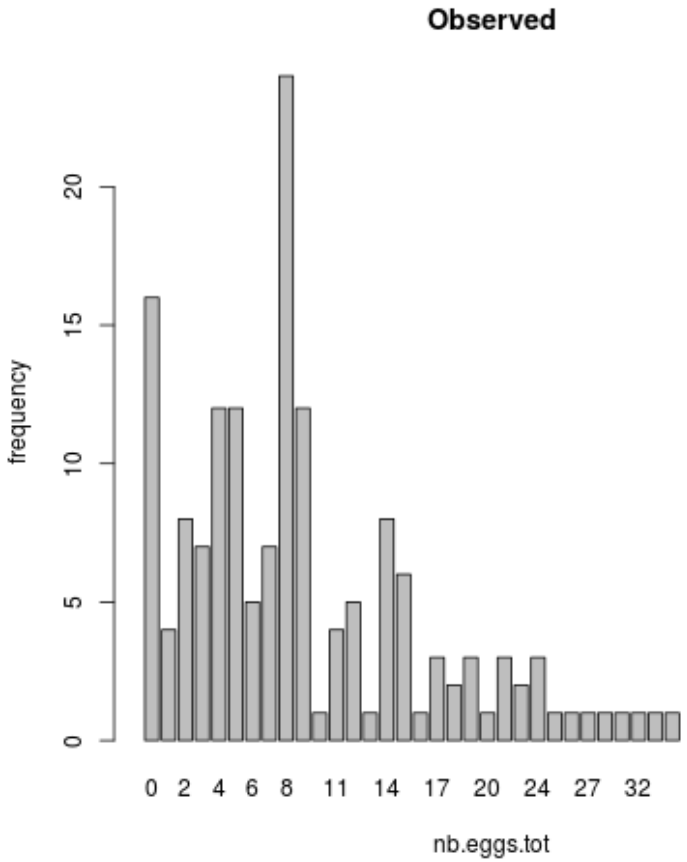
on the count scale: $\mu = e^{\beta_0 + \beta_1 X_1 + ... + \beta_k X_k} \in [0, +\infty[$



Example: number of worm eggs against time
roughly from Manon Grosmaire experiments

n measurements of size of progeny (=nb eggs) on 2 time points.





Is there eggs production?

```
glm0 <- glm(nb.eggs.tot ~ 1, data, family = "poisson")
```

The null model is:

$$log(\mu) = \beta_0$$

Example: number of worm eggs against time
Null model

Is there eggs production?

```
glm0 <- glm(nb.eggs.tot ~ 1, data, family = "poisson")
```

The null model is:

$$log(\mu) = \beta_0$$

```
> summary(glm0)

Call:
glm(formula = nb.eggs.tot ~ 1, family = "poisson", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3166  -1.9670  -0.4421   1.3548   8.3888

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.23178    0.02606   85.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1038.3  on 157  degrees of freedom
Residual deviance: 1038.3  on 157  degrees of freedom
AIC: 1598.9

Number of Fisher Scoring iterations: 5
```

Example: number of worm eggs against time

Null model: coefficient interpretation

If the model is true, asymptotically, estimators are gaussian.

```
> summary(glm0)

Call:
glm(formula = nb.eggs.tot ~ 1, family = "poisson", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3166  -1.9670  -0.4421   1.3548   8.3888

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.23178    0.02606   85.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1038.3  on 157  degrees of freedom
Residual deviance: 1038.3  on 157  degrees of freedom
AIC: 1598.9

Number of Fisher Scoring iterations: 5
```

- Recalling, $log(\mu) = \beta_0$, thus $\mu = e^{\beta_0}$.
- $IC_{95\%}(\beta_0) = \beta_0 \pm 1.96 \times \sigma_{\beta_0}$

Example: number of worm eggs against time

Null model: coefficient interpretation

If the model is true, asymptotically, estimators are gaussian.

```
> summary(glm0)

Call:
glm(formula = nb.eggs.tot ~ 1, family = "poisson", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3166  -1.9670  -0.4421   1.3548   8.3888

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.23178    0.02606   85.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1038.3  on 157  degrees of freedom
Residual deviance: 1038.3  on 157  degrees of freedom
AIC: 1598.9

Number of Fisher Scoring iterations: 5
```

- Recalling, $log(\mu) = \beta_0$, thus $\mu = e^{\beta_0}$.
- $IC_{95\%}(\beta_0) = \beta_0 \pm 1.96 \times \sigma_{\beta_0}$
- The mean number of eggs predicted is $e^{\beta_0} \simeq 9.3[8.9, 9.8]$ which is significant (Wald test, p-value <5%).

Example: number of worm eggs against time

Day effect model

Is there a change of eggs production according to day?

```
glm1 <- glm(nb.eggs.tot ~ day, data, family = "poisson")
```

The model is:

$$log(\mu) = \beta_0 + \beta_1 \times day$$

⚠ day is a factor: $day = 0$ for day 1, $day = 1$ for day 2).

Example: number of worm eggs against time
Day effect model

Is there a change of eggs production according to day?

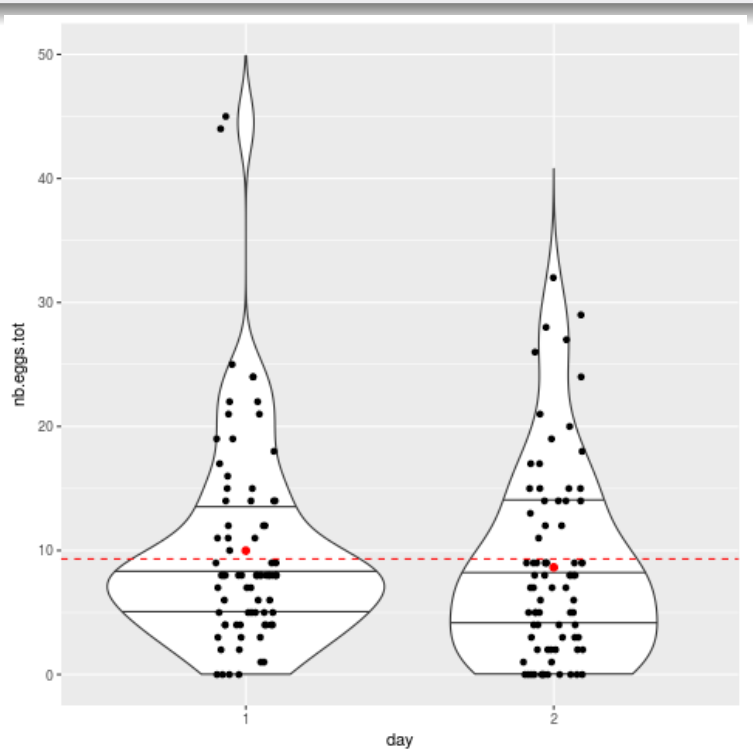
```
glm1 <- glm(nb.eggs.tot ~ day, data, family = "poisson")
```

The model is:

$$log(\mu) = \beta_0 + \beta_1 \times day$$

⚠ day is a factor: $day = 0$ for day 1, $day = 1$ for day 2).

```
> summary(glm1)

Call:
glm(formula = nb.eggs.tot ~ day, family = "poisson", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4693  -2.1574  -0.6517   1.1964   8.0905

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.30132    0.03560   64.64  < 2e-16 ***
day2        -0.14427    0.05226   -2.76  0.00577 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1038.3  on 157  degrees of freedom
Residual deviance: 1030.6  on 156  degrees of freedom
AIC: 1593.3

Number of Fisher Scoring iterations: 5
```



Example: number of worm eggs against time
Likelihood vs Deviance

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.

Example: number of worm eggs against time
Likelihood vs Deviance

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$
  - Null model: $L_{null} = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$
  - Null model: $L_{null} = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$
  - Effect of x_1: $L_{x_1} = \prod_{i=1}^{n} \frac{exp(\beta_0+\beta_1 x_{1i})^{y_i} e^{-exp(\beta_0+\beta_1 x_{1i})}}{y_i!}$

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$
  - Null model: $L_{null} = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$
  - Effect of x_1: $L_{x_1} = \prod_{i=1}^{n} \frac{exp(\beta_0+\beta_1 x_{1i})^{y_i} e^{-exp(\beta_0+\beta_1 x_{1i})}}{y_i!}$
- The **deviance** is a variation of **log**-likelihood, LL.

Example: number of worm eggs against time

Likelihood vs Deviance

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$
  - Null model: $L_{null} = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$
  - Effect of x_1: $L_{x_1} = \prod_{i=1}^{n} \frac{exp(\beta_0+\beta_1 x_{1i})^{y_i} e^{-exp(\beta_0+\beta_1 x_{1i})}}{y_i!}$
- The **deviance** is a variation of **log**-likelihood, LL.
  - $D_{null} = -2(LL_{null} - LL_{sat}) = -2LL_{null}$

Example: number of worm eggs against time

Likelihood vs Deviance

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$
  - Null model: $L_{null} = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$
  - Effect of x_1: $L_{x_1} = \prod_{i=1}^{n} \frac{exp(\beta_0+\beta_1 x_{1i})^{y_i} e^{-exp(\beta_0+\beta_1 x_{1i})}}{y_i!}$
- The **deviance** is a variation of **log**-likelihood, LL.
  - $D_{null} = -2(LL_{null} - LL_{sat}) = -2LL_{null}$
  - $D_{x_1} = -2(LL_{x_1} - LL_{sat}) = -2LL_{x_1}$

Example: number of worm eggs against time

Likelihood vs Deviance

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$
  - Null model: $L_{null} = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$
  - Effect of x_1: $L_{x_1} = \prod_{i=1}^{n} \frac{exp(\beta_0+\beta_1 x_{1i})^{y_i} e^{-exp(\beta_0+\beta_1 x_{1i})}}{y_i!}$
- The **deviance** is a variation of **log**-likelihood, LL.
  - $D_{null} = -2(LL_{null} - LL_{sat}) = -2LL_{null}$
  - $D_{x_1} = -2(LL_{x_1} - LL_{sat}) = -2LL_{x_1}$
- Models comparisons:
  $D_{effect\ x_1} = D_{null} - D_{x_1}$ $\hookrightarrow \chi^2((n - p_{null}) - (n - p_{x_1}))$

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$
  - Null model: $L_{null} = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$
  - Effect of x_1: $L_{x_1} = \prod_{i=1}^{n} \frac{exp(\beta_0+\beta_1 x_{1i})^{y_i} e^{-exp(\beta_0+\beta_1 x_{1i})}}{y_i!}$
- The **deviance** is a variation of **log**-likelihood, LL.
  - $D_{null} = -2(LL_{null} - LL_{sat}) = -2LL_{null}$
  - $D_{x_1} = -2(LL_{x_1} - LL_{sat}) = -2LL_{x_1}$
- Models comparisons:

$$\begin{aligned} D_{effect\ x_1} &= D_{null} - D_{x_1} && \hookrightarrow \chi^2((n-p_{null}) - (n-p_{x_1})) \\ &= -2(LL_{null} - LL_{x_1}) && \hookrightarrow \chi^2(p_{x_1} - p_{null}) \end{aligned}$$

- **Likelihood**, L, is $\prod_{i=1}^{n} P(Y_i = y_i)$.
  - Saturated model: $L_{sat} = 1$
  - Null model: $L_{null} = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$
  - Effect of x_1: $L_{x_1} = \prod_{i=1}^{n} \frac{exp(\beta_0+\beta_1 x_{1i})^{y_i} e^{-exp(\beta_0+\beta_1 x_{1i})}}{y_i!}$
- The **deviance** is a variation of **log**-likelihood, LL.
  - $D_{null} = -2(LL_{null} - LL_{sat}) = -2LL_{null}$
  - $D_{x_1} = -2(LL_{x_1} - LL_{sat}) = -2LL_{x_1}$
- Models comparisons:

$$\begin{aligned} D_{effect\ x_1} &= D_{null} - D_{x_1} && \hookrightarrow \chi^2((n-p_{null}) - (n-p_{x_1})) \\ &= -2(LL_{null} - LL_{x_1}) && \hookrightarrow \chi^2(p_{x_1} - p_{null}) \end{aligned}$$

If $\chi^2_{obs} < \chi^2_{th}$, the 2 models are not statistically different and you should choose the more parsimonious.

- On linear predictor scale, $log(\mu) = \beta_0 + \beta_1 \times day$

Example: number of worm eggs against time

Day effect model: coefficients interpretation

- On linear predictor scale, $log(\mu) = \beta_0 + \beta_1 \times day$
- On counts scale, $\mu = e^{\beta_0} e^{\beta_1 \times day}$
  - For day 1: $day = 0$, thus $\mu_{day_1} = e^{\beta_0}$
  - For day 2: $day = 1$, thus $\mu_{day_2} = e^{\beta_0} e^{\beta_1}$

- On linear predictor scale, $log(\mu) = \beta_0 + \beta_1 \times day$
- On counts scale, $\mu = e^{\beta_0} e^{\beta_1 \times day}$
  - For day 1: $day = 0$, thus $\mu_{day_1} = e^{\beta_0}$
  - For day 2: $day = 1$, thus $\mu_{day_2} = e^{\beta_0} e^{\beta_1}$

$e^{\beta_1} = \frac{\mu_{day_2}}{\mu_{day_1}} = 0.9[0.78, 0.96]$ indicates effect of ageing is significantly deleterious for eggs production (Wald test, p-value <5%).

Example: number of worm eggs against time

Day effect model: coefficients interpretation

- On linear predictor scale, $log(\mu) = \beta_0 + \beta_1 \times day$
- On counts scale, $\mu = e^{\beta_0} e^{\beta_1 \times day}$
  - For day 1: $day = 0$, thus $\mu_{day_1} = e^{\beta_0}$
  - For day 2: $day = 1$, thus $\mu_{day_2} = e^{\beta_0} e^{\beta_1}$

  $e^{\beta_1} = \frac{\mu_{day_2}}{\mu_{day_1}} = 0.9[0.78, 0.96]$ indicates effect of ageing is significantly deleterious for eggs production (Wald test, p-value <5%).
- Is it though relevant to add a day effect?

```
> anova(glm0, glm1, test = "Chisq")
Analysis of Deviance Table

Model 1: nb.eggs.tot ~ 1
Model 2: nb.eggs.tot ~ day
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       157     1038.3
2       156     1030.6  1   7.6398  0.00571 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example: number of worm eggs against time

Day effect model: coefficients interpretation

- On linear predictor scale, $log(\mu) = \beta_0 + \beta_1 \times day$
- On counts scale, $\mu = e^{\beta_0} e^{\beta_1 \times day}$
  - For day 1: $day = 0$, thus $\mu_{day_1} = e^{\beta_0}$
  - For day 2: $day = 1$, thus $\mu_{day_2} = e^{\beta_0} e^{\beta_1}$

  $e^{\beta_1} = \frac{\mu_{day_2}}{\mu_{day_1}} = 0.9[0.78, 0.96]$ indicates effect of ageing is significantly deleterious for eggs production (Wald test, p-value <5%).
- Is it though relevant to add a day effect?

```
> anova(glm0, glm1, test = "Chisq")
Analysis of Deviance Table

Model 1: nb.eggs.tot ~ 1
Model 2: nb.eggs.tot ~ day
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       157     1038.3
2       156     1030.6  1   7.6398  0.00571 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$\chi^2_{obs} < \chi^2_{th}$, the day effect model explains significantly better the variability of the data (`1-pchisq(7.6398, df = 1)`).

Example: number of worm eggs against time
Zoom on model assumption

- Do we have $E(Y) = Var(Y)$?

- Do we have E(Y) = Var(Y)?

Residual variance is estimated by $\frac{1}{n-p}\sum_{i=1}^{n}(y_i-\mu_i)^2$.

If Poisson law and model are adapted, $\frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\mu_i} \sim 1$

- Do we have E(Y) = Var(Y)?

Residual variance is estimated by $\frac{1}{n-p}\sum_{i=1}^{n}(y_i-\mu_i)^2$.

If Poisson law and model are adapted, $\frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\mu_i} \sim 1$

or here it equals 6.98, indicating $E(Y) < Var(Y)$.
Tests are not reliable and this is not the best model; it causes sd error to be deflated which could lead to significant predictor whereas it is not.

To deal with over-dispersion, knowing $E(Y) = \mu$

- quasiPoisson: $Var(Y) = \phi\mu$

```
glm1b <- glm(nb.eggs.tot ~ day, data, family = "quasipoisson")
```

To deal with over-dispersion, knowing $E(Y) = \mu$

- quasiPoisson: $Var(Y) = \phi\mu$

```
glm1b <- glm(nb.eggs.tot ~ day, data, family = "quasipoisson")
```

- Negative binomial: $Y \hookrightarrow Pois(\Theta \times \mu)$ and $\Theta \hookrightarrow Gamma(\alpha, \alpha)$ with $E(\Theta) = 1$
  $Var(Y) = \mu + \alpha \times \mu^2$
  The link function becomes: $log(\frac{\alpha\mu}{1+\alpha\mu})$.

```
glm2 <- MASS::glm.nb(nb.eggs.tot ~ day, data)
```

Example: number of worm eggs against time
Alternatives

To deal with over-dispersion, knowing $E(Y) = \mu$

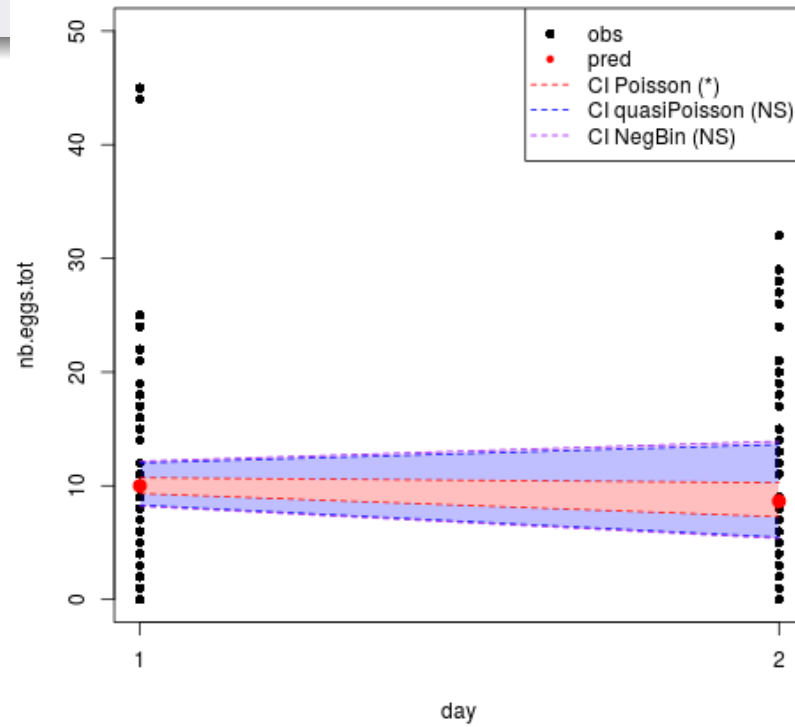
- quasiPoisson: $Var(Y) = \phi\mu$

```
glm1b <- glm(nb.eggs.tot ~ day, data, family = "quasipoisson")
```

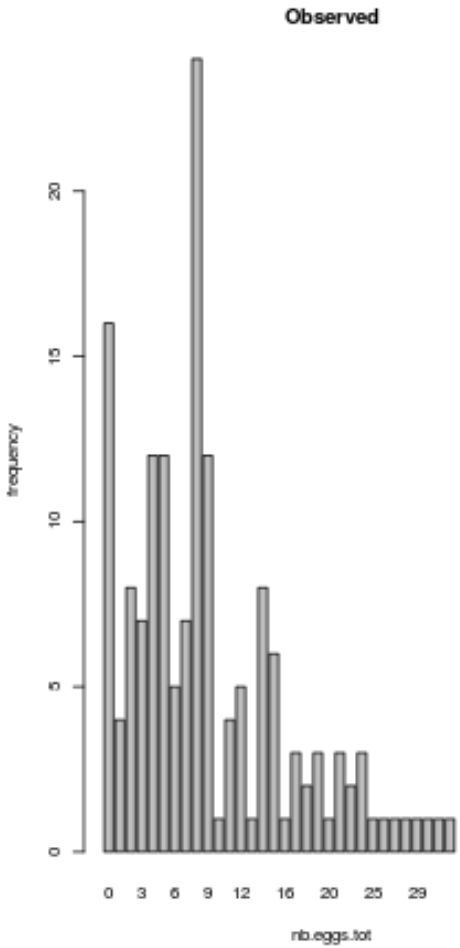
- Negative binomial: $Y \hookrightarrow Pois(\Theta \times \mu)$ and $\Theta \hookrightarrow Gamma(\alpha, \alpha)$ with $E(\Theta) = 1$
  $Var(Y) = \mu + \alpha \times \mu^2$
  The link function becomes: $log(\frac{\alpha\mu}{1+\alpha\mu})$.

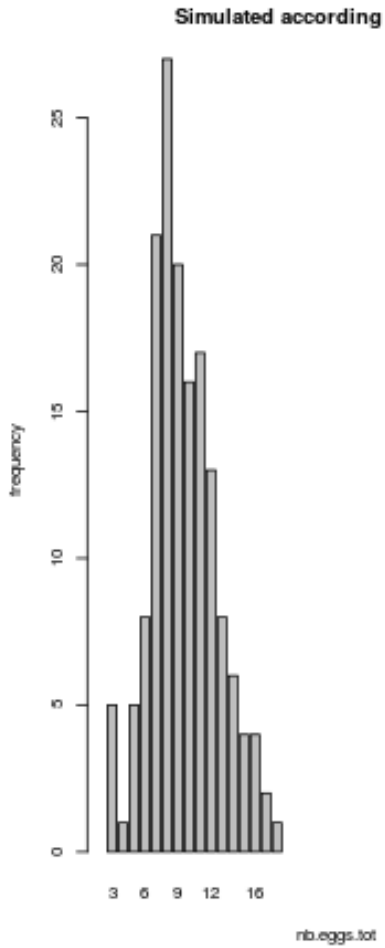
```
glm2 <- MASS::glm.nb(nb.eggs.tot ~ day, data)
```

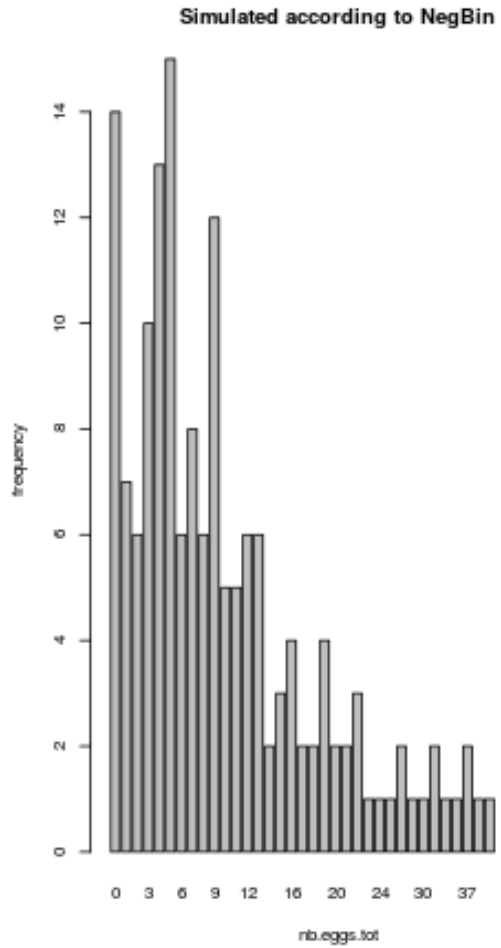
```
> summary(glm1b)

Call:
glm(formula = nb.eggs.tot ~ day, family = "quasipoisson", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4693  -2.1574  -0.6517   1.1964   8.0905

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.30132    0.09407  24.464   <2e-16 ***
day2        -0.14427    0.13810  -1.045    0.298
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 6.981896)

    Null deviance: 1038.3  on 157  degrees of freedom
Residual deviance: 1030.6  on 156  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```



Example: number of worm eggs against time
Alternatives

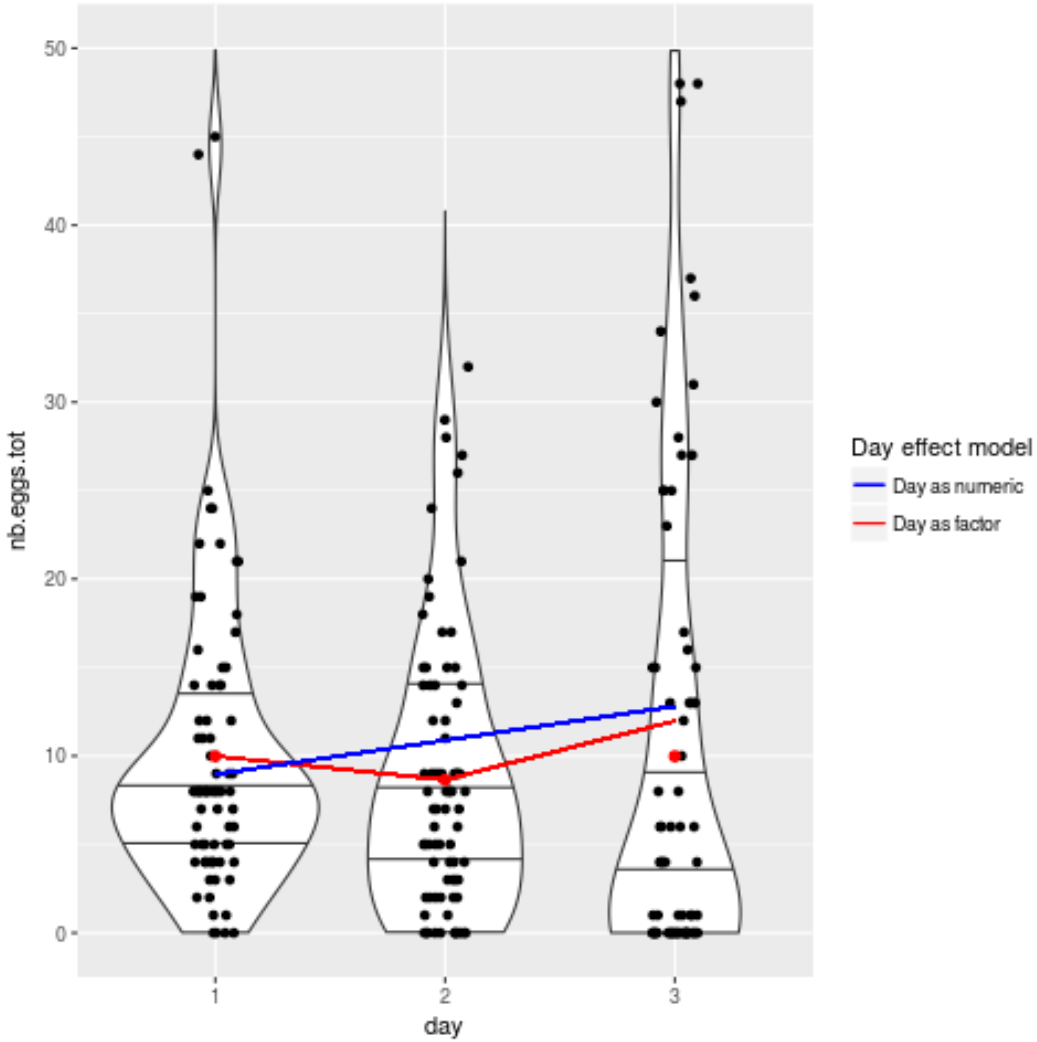






Example: number of worm eggs against time
Last but not least 1/2

- Why to insist on type of data?



Example: number of worm eggs against time
Last but not least 2/2

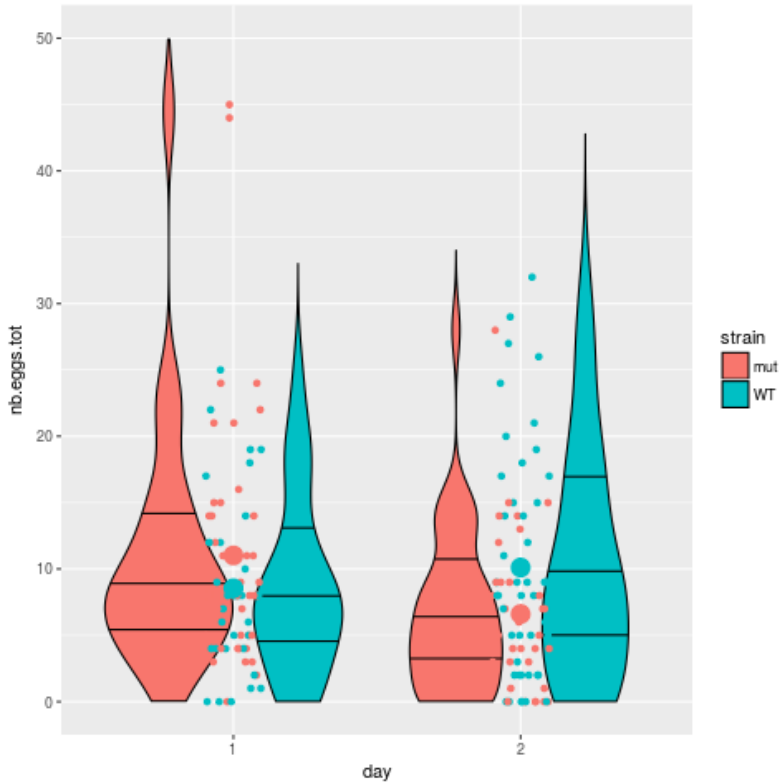
- Quick display of interaction: add the *strain* effect.

Without interaction:

```
glm1ds <- glm(nb.eggs.tot ~ day+strain, data,
        family = "poisson")
```

$$lm(\mu) = \beta_0 + \beta_1 \times day + \beta_2 \times strain$$

$$\mu_{day_2/WT} = e^{\beta 0}\, e^{\beta 1}\, e^{\beta 2}$$



Example: number of worm eggs against time
Last but not least 2/2

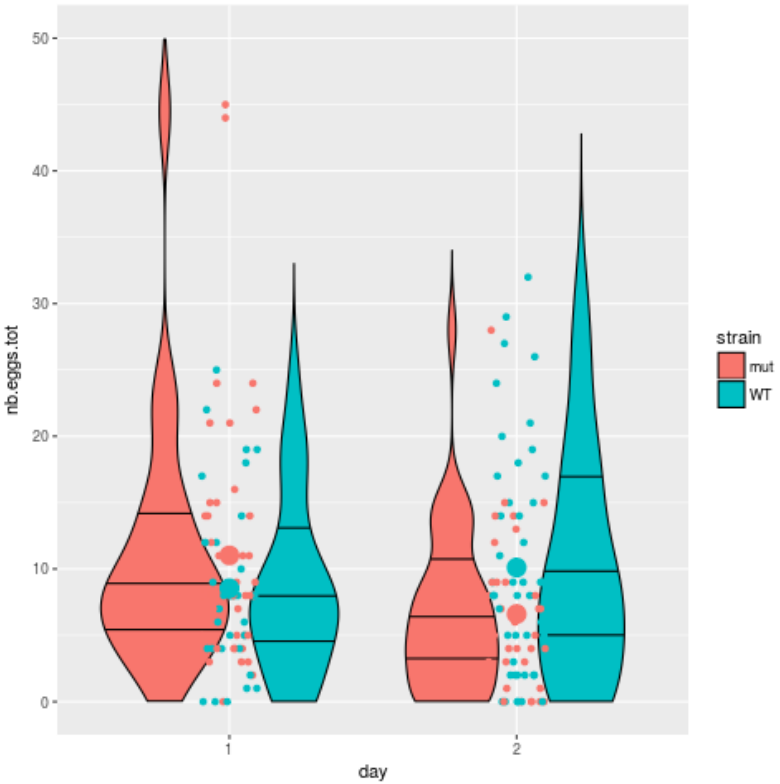
- Quick display of interaction: add the *strain* effect.

Without interaction:

```
glm1ds <- glm(nb.eggs.tot ~ day+strain, data,
        family = "poisson")
```

$$lm(\mu) = \beta_0 + \beta_1 \times day + \beta_2 \times strain$$

$$\mu_{day_2/WT} = e^{\beta 0}\, e^{\beta 1}\, e^{\beta 2}$$



With interaction:

```
glm1dsI <- glm(nb.eggs.tot ~ day*strain, data, family = "poisson")
```

Example: number of worm eggs against time
Last but not least 2/2

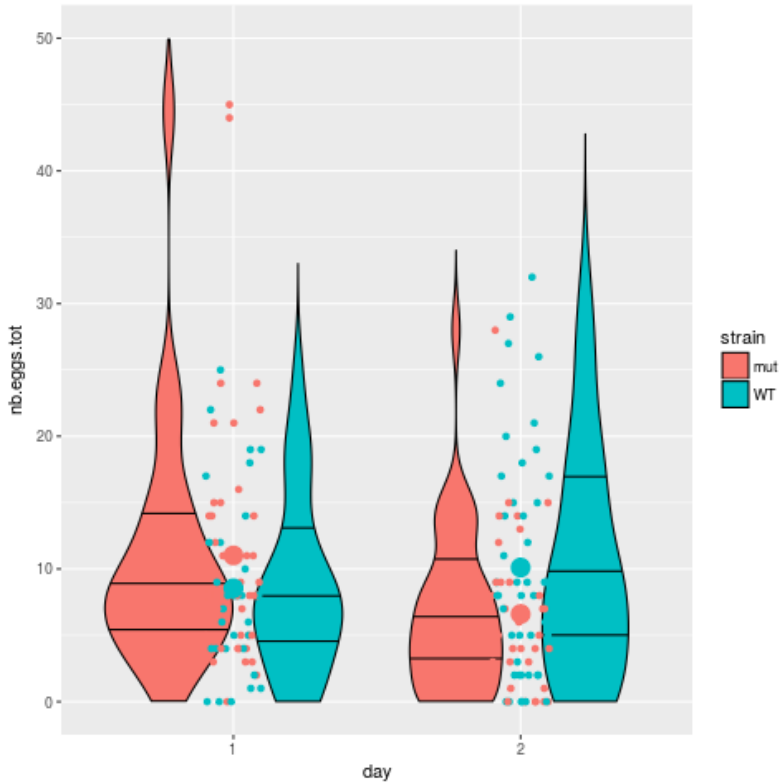
- Quick display of interaction: add the *strain* effect.

Without interaction:

```
glm1ds <- glm(nb.eggs.tot ~ day+strain, data,
        family = "poisson")
```

$$lm(\mu) = \beta_0 + \beta_1 \times day + \beta_2 \times strain$$

$$\mu_{day_2/WT} = e^{\beta 0}\, e^{\beta 1}\, e^{\beta 2}$$



With interaction:

```
glm1dsI <- glm(nb.eggs.tot ~ day*strain, data, family = "poisson")
```

$$lm(\mu) = \beta_0 + \beta_1 \times day + \beta_2 \times strain + \beta_3 \times day \times strain$$

$$\mu_{day_2/WT} = e^{\beta 0}\, e^{\beta 1}\, e^{\beta 2}\, e^{\beta 3}$$

Example: number of worm eggs against time
Last but not least 2/2

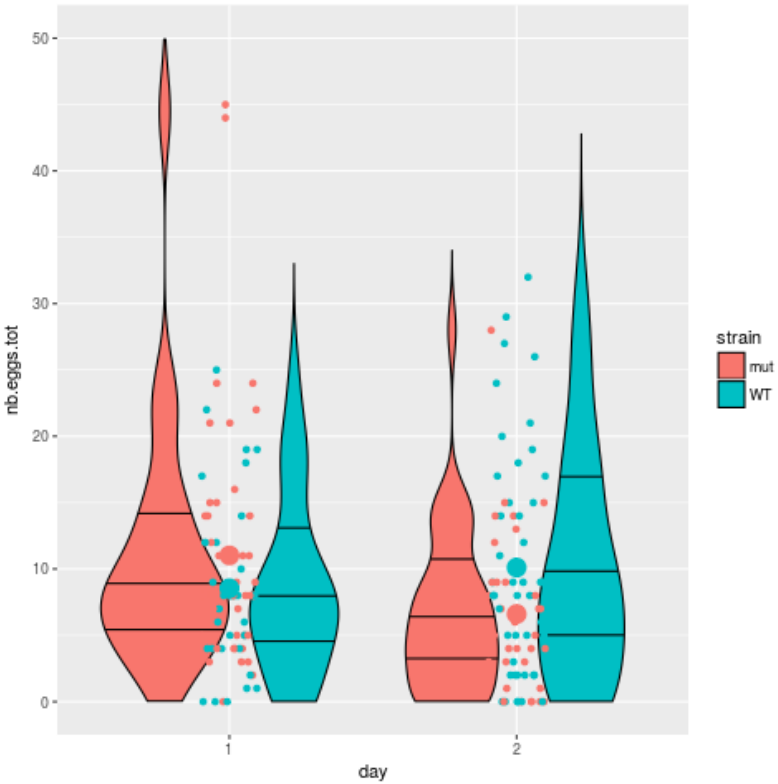
- Quick display of interaction: add the *strain* effect.

Without interaction:

```
glm1ds <- glm(nb.eggs.tot ~ day+strain, data,
        family = "poisson")
```

$lm(\mu) = \beta_0 + \beta_1 \times day + \beta_2 \times strain$

$\mu_{day_2/WT} = e^{\beta 0}\, e^{\beta 1}\, e^{\beta 2}$



With interaction:

```
glm1dsI <- glm(nb.eggs.tot ~ day*strain, data, family = "poisson")
```

$lm(\mu) = \beta_0 + \beta_1 \times day + \beta_2 \times strain + \beta_3 \times day \times strain$

$\mu_{day_2/WT} = e^{\beta 0}\, e^{\beta 1}\, e^{\beta 2}\, e^{\beta 3}$

```
> summary(glm1dsI)

Call:
glm(formula = nb.eggs.tot ~ day * strain, family = "poisson",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6950  -1.9715  -0.6709   1.1131   7.6586

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.39987    0.04441  54.037  < 2e-16 ***
day2         -0.51188    0.08099  -6.320 2.61e-10 ***
strainWT     -0.25447    0.07429  -3.426 0.000614 ***
day2:strainWT  0.67988    0.11071   6.141 8.19e-10 ***
```

Example: number of worm eggs against time
Last but not least 2/2

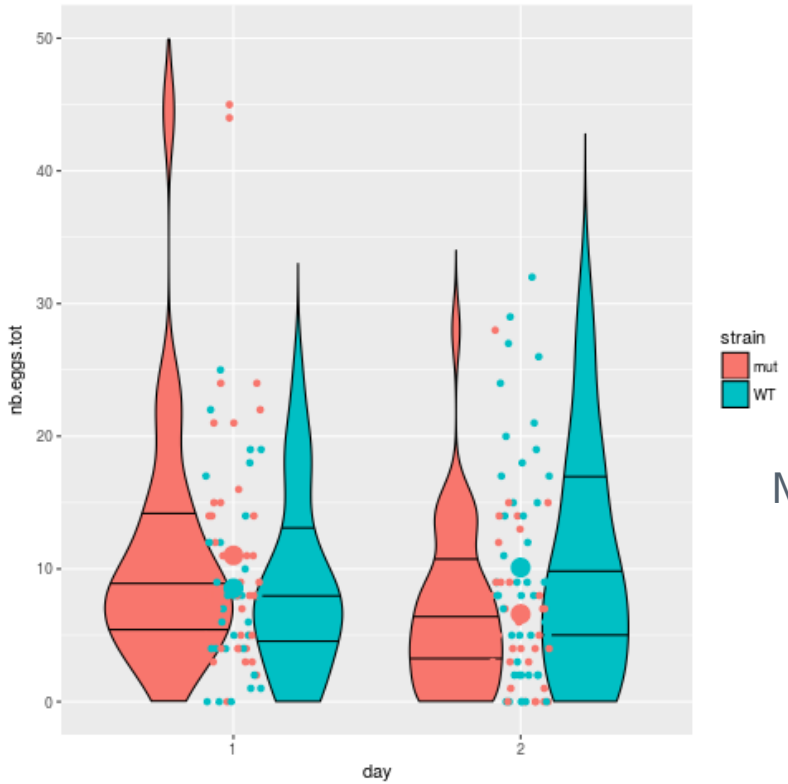
- Quick display of interaction: add the *strain* effect.

Without interaction:

```
glm1ds <- glm(nb.eggs.tot ~ day+strain, data,
        family = "poisson")
```

$lm(\mu) = \beta_0 + \beta_1 \times day + \beta_2 \times strain$

$\mu_{day_2/WT} = e^{\beta 0}\, e^{\beta 1}\, e^{\beta 2}$



With interaction:

```
glm1dsI <- glm(nb.eggs.tot ~ day*strain, data, family = "poisson")
```

$lm(\mu) = \beta_0 + \beta_1 \times day + \beta_2 \times strain + \beta_3 \times day \times strain$

$\mu_{day_2/WT} = e^{\beta 0}\, e^{\beta 1}\, e^{\beta 2}\, e^{\beta 3}$

```
> summary(glm1dsI)

Call:
glm(formula = nb.eggs.tot ~ day * strain, family = "poisson",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6950  -1.9715  -0.6709   1.1131   7.6586

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.39987    0.04441  54.037  < 2e-16 ***
day2         -0.51188    0.08099  -6.320 2.61e-10 ***
strainWT     -0.25447    0.07429  -3.426 0.000614 ***
day2:strainWT  0.67988    0.11071   6.141 8.19e-10 ***
```

Models comparisons:

```
> anova(glm1ds, glm1dsI, test = "Chisq")
Analysis of Deviance Table

Model 1: nb.eggs.tot ~ day + strain
Model 2: nb.eggs.tot ~ day * strain
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       155    1029.53
2       154     990.53  1   39.001 4.236e-10 ***
```

Binomial distribution

Proportion/% data

Y is the number of times an event occur on n tries.

- $Y \hookrightarrow Binom(n,p)$ with $P(Y=k) = C_n^k (1-p)^{(n-k)}$

Binomial distribution

Proportion/% data

Y is the number of times an event occur on n tries.

- $Y \hookrightarrow Binom(n,p)$ with $P(Y=k) = C_n^k (1-p)^{(n-k)}$
- We are more interested on the frequency of the event $\mu = \frac{Y}{n}$
- $E(\frac{Y}{n}) = p$ and $var(\frac{Y}{n}) = \frac{p(1-p)}{n}$

Binomial distribution

Proportion/% data

Y is the number of times an event occur on n tries.

- $Y \hookrightarrow Binom(n,p)$ with $P(Y=k) = C_n^k (1-p)^{(n-k)}$
- We are more interested on the frequency of the event $\mu = \frac{Y}{n}$
- $E(\frac{Y}{n}) = p$ and $var(\frac{Y}{n}) = \frac{p(1-p)}{n}$
- The link function is the logit function, the model is:

$$logit(\mu) = log(\frac{\mu}{1-\mu}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \in]-\infty, +\infty[$$

Binomial distribution

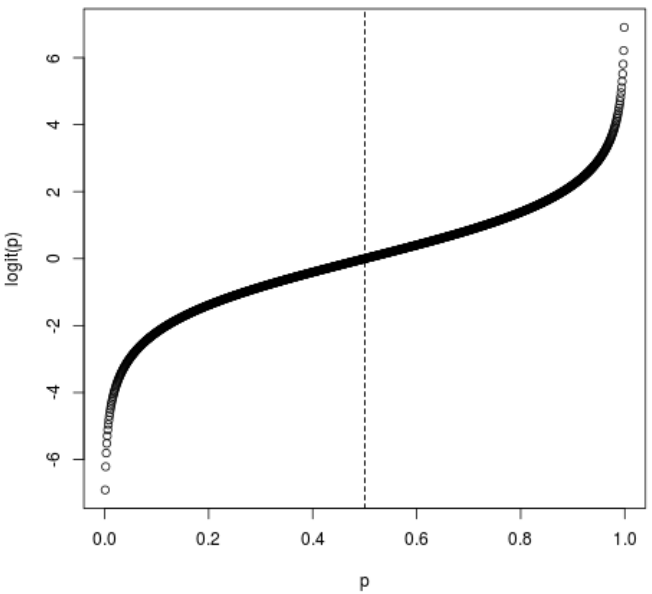
Proportion/% data

Y is the number of times an event occur on n tries.

- $Y \hookrightarrow Binom(n,p)$ with $P(Y=k) = C_n^k (1-p)^{(n-k)}$
- We are more interested on the frequency of the event $\mu = \frac{Y}{n}$
- $E(\frac{Y}{n}) = p$ and $var(\frac{Y}{n}) = \frac{p(1-p)}{n}$
- The link function is the logit function, the model is:

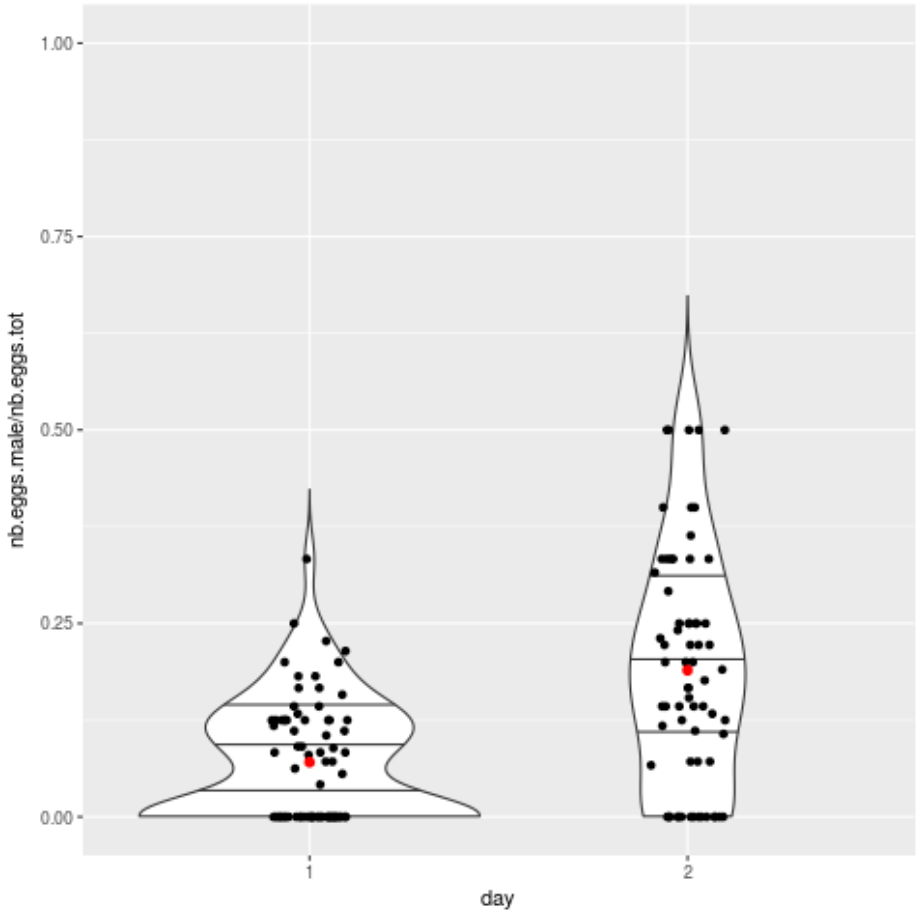
$$logit(\mu) = log(\frac{\mu}{1-\mu}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \in]-\infty, +\infty[$$

on the proportion scale: $\mu = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+...+\beta_k X_k)}} \in [0,1]$



Example: proportion of male worm eggs against time
from Manon Grosmaire experiments

n measurements of male progeny (=nb eggs) on 2 time points.



Is there male eggs production?

```
bglm0 <- glm(cbind(nb.eggs.male, nb.eggs.tot-nb.eggs.male) ~ 1,
data = data, family = "binomial")
```

The null model is:

$$log(\frac{\mu}{1-\mu}) = log(\frac{p_{male}}{1-p_{male}}) = log(odd) = \beta_0 \Leftrightarrow odd = e^{\beta_0}$$

Is there male eggs production?

```
bglm0 <- glm(cbind(nb.eggs.male, nb.eggs.tot-nb.eggs.male) ~ 1,
data = data, family = "binomial")
```

The null model is:

$$log(\frac{\mu}{1-\mu}) = log(\frac{p_{male}}{1-p_{male}}) = log(odd) = \beta_0 \Leftrightarrow odd = e^{\beta_0}$$

On 4 eggs, an odds equal to 3 indicates that 3 eggs against 1 will be male.
The more the odds, the more the probability.

Example: proportion of male worm eggs against time

Null model: coefficient interpretation

If the model is true, asymptotically, estimators are gaussian.

```
> summary(bglm0)

Call:
glm(formula = cbind(nb.eggs.male, nb.eggs.tot - nb.eggs.male) ~
    1, family = "binomial", data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.04765  -1.01609  -0.04592   0.43340   2.20961

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.89712    0.07739  -24.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 159.62  on 141  degrees of freedom
Residual deviance: 159.62  on 141  degrees of freedom
AIC: 369.72

Number of Fisher Scoring iterations: 4
```

- Recalling, $odd = e^{\beta_0}$.
- $IC_{95\%}(\beta_0) = \beta_0 \pm 1.96 \times \sigma_{\beta_0}$

Example: proportion of male worm eggs against time
Null model: coefficient interpretation

If the model is true, asymptotically, estimators are gaussian.

```
> summary(bglm0)

Call:
glm(formula = cbind(nb.eggs.male, nb.eggs.tot - nb.eggs.male) ~
    1, family = "binomial", data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.04765  -1.01609  -0.04592   0.43340   2.20961

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.89712    0.07739  -24.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 159.62  on 141  degrees of freedom
Residual deviance: 159.62  on 141  degrees of freedom
AIC: 369.72

Number of Fisher Scoring iterations: 4
```

- Recalling, $odd = e^{\beta_0}$.
- $IC_{95\%}(\beta_0) = \beta_0 \pm 1.96 \times \sigma_{\beta_0}$
- The odd of male birth predicted is $e^{\beta_0} \simeq 0.15[0.13, 0.17]$ which is significant (Wald test, p-value <5%).

Example: proportion of male worm eggs against time
Day effect model

Is there a change of male eggs production according to day?

```
bglm1 <- glm(cbind(nb.eggs.male, nb.eggs.tot-nb.eggs.male)
~ day, data = data, family = "binomial")
```

The model is:

$$log(odd) = \beta_0 + \beta_1 \times day$$

⚠ *day* is a factor: $day = 0$ for day 1, $day = 1$ for day 2.

Example: proportion of male worm eggs against time

Day effect model: coefficients interpretation

- On linear predictor scale, $log(odd) = \beta_0 + \beta_1 \times day$
  - For day 1: $day = 0$, thus $log(odd_{day_1}) = \beta_0$

Example: proportion of male worm eggs against time

Day effect model: coefficients interpretation

- On linear predictor scale, $log(odd) = \beta_0 + \beta_1 \times day$
  - For day 1: $day = 0$, thus $log(odd_{day_1}) = \beta_0$
  - For day 2: $day = 1$, thus $log(odd_{day_2}) = \beta_0 + \beta_1$

Example: proportion of male worm eggs against tim

Day effect model: coefficients interpretation

- On linear predictor scale, $log(odd) = \beta_0 + \beta_1 \times day$
  - For day 1: $day = 0$, thus $log(odd_{day_1}) = \beta_0$
  - For day 2: $day = 1$, thus $log(odd_{day_2}) = \beta_0 + \beta_1$
- Towards Odds Ratio:
  - $log(\frac{odd_{day_2}}{odd_{day_1}}) = log(OR_{days}) = \beta_1$

Example: proportion of male worm eggs against tim

Day effect model: coefficients interpretation

- On linear predictor scale, $log(odd) = \beta_0 + \beta_1 \times day$
  - For day 1: $day = 0$, thus $log(odd_{day_1}) = \beta_0$
  - For day 2: $day = 1$, thus $log(odd_{day_2}) = \beta_0 + \beta_1$
- Towards Odds Ratio:
  - $log(\frac{odd_{day_2}}{odd_{day_1}}) = log(OR_{days}) = \beta_1$
  - $OR_{days} = e^{\beta_1} = \times$ odd from day 1 to 2

Example: proportion of male worm eggs against time

Day effect model: coefficients interpretation

- On linear predictor scale, $log(odd) = \beta_0 + \beta_1 \times day$
  - For day 1: $day = 0$, thus $log(odd_{day_1}) = \beta_0$
  - For day 2: $day = 1$, thus $log(odd_{day_2}) = \beta_0 + \beta_1$
- Towards Odds Ratio:
  - $log(\frac{odd_{day_2}}{odd_{day_1}}) = log(OR_{days}) = \beta_1$
  - $OR_{days} = e^{\beta_1} =$ × odd from day 1 to 2
  - $OR_{days} = 2.2[1.6, 3.0] > 1$ suggests ageing tends to favor chances to get males (Wald test, p-value <5%).

```
> summary(bglm1)

Call:
glm(formula = cbind(nb.eggs.male, nb.eggs.tot - nb.eggs.male) ~
    day, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4297  -0.8809   0.0000   0.3562   2.0473

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3294     0.1252 -18.605  < 2e-16 ***
day2          0.8037     0.1602   5.017 5.23e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 159.62  on 141  degrees of freedom
Residual deviance: 133.45  on 140  degrees of freedom
AIC: 345.55

Number of Fisher Scoring iterations: 4
```

- Is it though relevant to add a day effect?

- Is it though relevant to add a day effect?

```
> anova(glm0, glm1, test = "Chisq")
Analysis of Deviance Table

Model 1: nb.eggs.tot ~ 1
Model 2: nb.eggs.tot ~ day
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       157     1038.3
2       156     1030.6  1   7.6398  0.00571 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Is it though relevant to add a day effect?

```
> anova(glm0, glm1, test = "Chisq")
Analysis of Deviance Table

Model 1: nb.eggs.tot ~ 1
Model 2: nb.eggs.tot ~ day
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       157     1038.3
2       156     1030.6  1   7.6398  0.00571 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$\chi^2_{obs} < \chi^2_{th}$, the day effect model explains significantly better the variability of the data (`1-pchisq(7.6398, df = 1)`).

The session is finished!

Thanks for coming!

- The next session will be on Machine Learning (co-clustering analysis) by Margot Selosse, PhD student, from Lyon II university, *Foodle to come*.
- A quick reminder about our Slack tchat.
- Any topics you want to discuss?