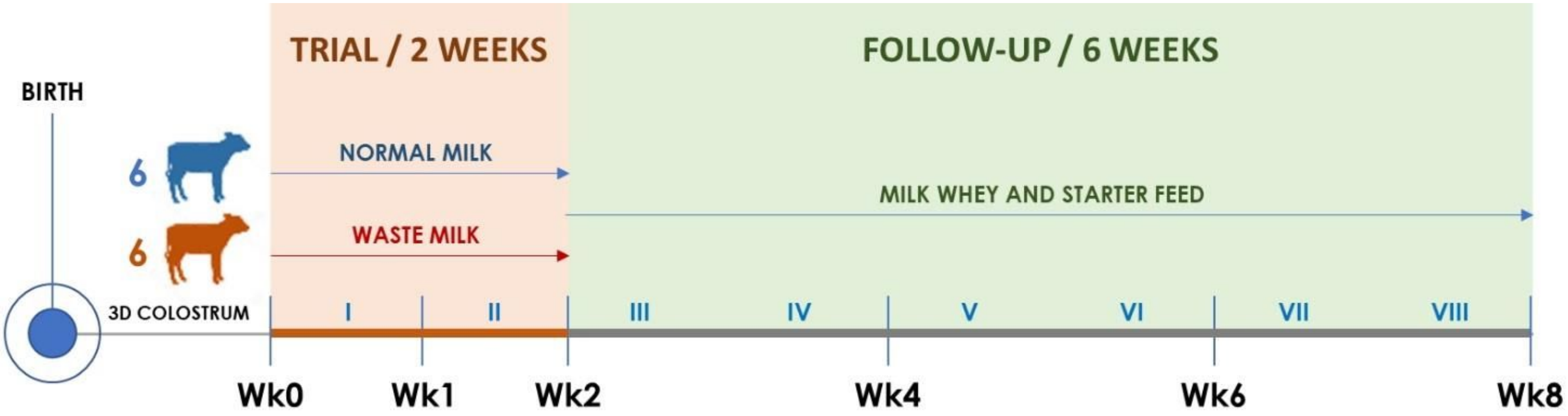
The basic experimental setting: treatments and timepoints

Filippo Biscarini

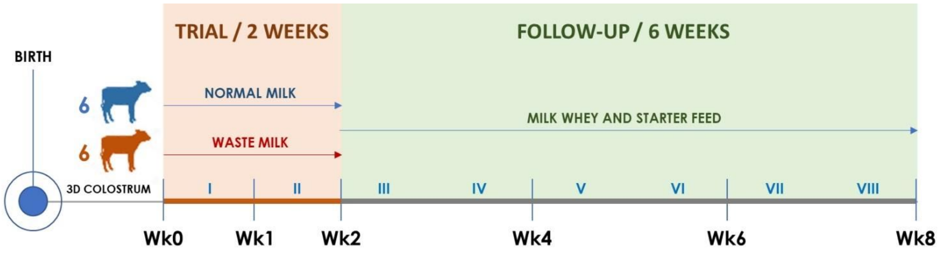
(Biostatistician, bioinformatician and quantitative geneticist) **CNR**, Milan (Italy)

Treatments and timepoints



From: Penati et al. 2021 (Front. Vet. Sci., 08 July 2021)

Treatments and timepoints



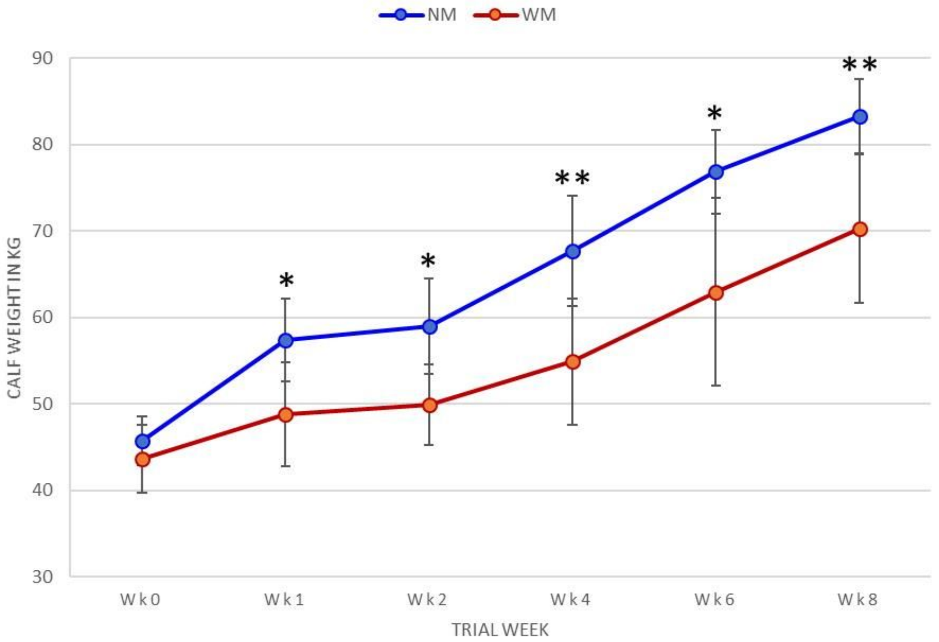
From: Penati et al. 2021 (Front. Vet. Sci., 08 July 2021)

- treatments: groups
- timepoints: longitudinal (time) component

- the effect of the treatment(s) is not expected at T0
- we may observe the effect of the treatment later in time, progressively stronger
- the effect may appear and then subside
- there may be no effect of the treatment
- etc.

Treatments and timepoints

Effect of treatment apparent from after one week, then remains constant or intensifies



* p <0.05 ** p <0.005

From: Penati et al. 2021 (Front. Vet. Sci., 08 July 2021)

Treatments and timepoints

- Is there an effect of treatment on the response variable?
  - e.g. drug vs placebo for disease response
- How strong is the effect of treatment?
  - e.g. how fast / how much does the drug reduce symptoms vs placebo?
- How accurately can we predict future/unobserved response variables?
  - e.g. date of remission based on treatment; fish growth based on diet; etc.
- Does the effect of treatment change with time?
  - e.g. lag of drug to combat disease; faster effect at first, decreasing returns with time; reversed effect of antimicrobial if continued for too long; etc.

Treatments and timepoints

- Is there an effect of treatment on the response variable?
  - e.g. drug vs placebo for disease response
- How strong is the effect of treatment?
  - e.g. how fast / how much does the drug reduce symptoms vs placebo?
- How accurately can we predict future/unobserved response variables?
  - e.g. date of remission based on treatment; fish growth based on diet; etc.
- Does the effect of treatment change with time?
  - e.g. lag of drug to combat disease; faster effect at first, decreasing returns with time; reversed effect of antimicrobial if continued for too long; etc.

Q: can you think of additional examples from your own research? Let's discuss them!

Simple linear regression

$$y = \beta_0 + \beta_1 x + e$$

Q: are you familiar with this equation?

Simple linear regression

$$y = \beta_0 + \beta_1 x + e$$

$$\text{response} = \beta_0 + \beta_1 \text{treatment} + e$$

Simple linear regression

$$y = \beta_0 + \beta_1 x + e$$

Q: how can we estimate the model coefficients (b0, b1)?

Least squares

Minimise the residuals → Q: how do we calculate the residuals of the model?
→ Q: what happens to the sum of the residuals?
→ Q: what do we do then?

Least squares

Minimise the residuals → Q: how do we calculate the residuals of the model?
→ Q: what happens to the sum of the residuals?
→ Q: what do we do then?

residuals = (y - (b0 + b1x))
sum(y - (b0 + b1x)) = 0

$$\mathrm{RSS} = \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i))^2$$

to be <u>minimised</u>

Least squares: the normal equations

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

We can use a more compact matrix notation:

y = Xb + e

where **y**, **b**, and **e** are vectors of responses, coefficients and residuals; **X** is the design matrix

e.g. **b = [b0, b1]**

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

(this is generalizable to multiple linear regression)

Estimating coefficients: alternatives

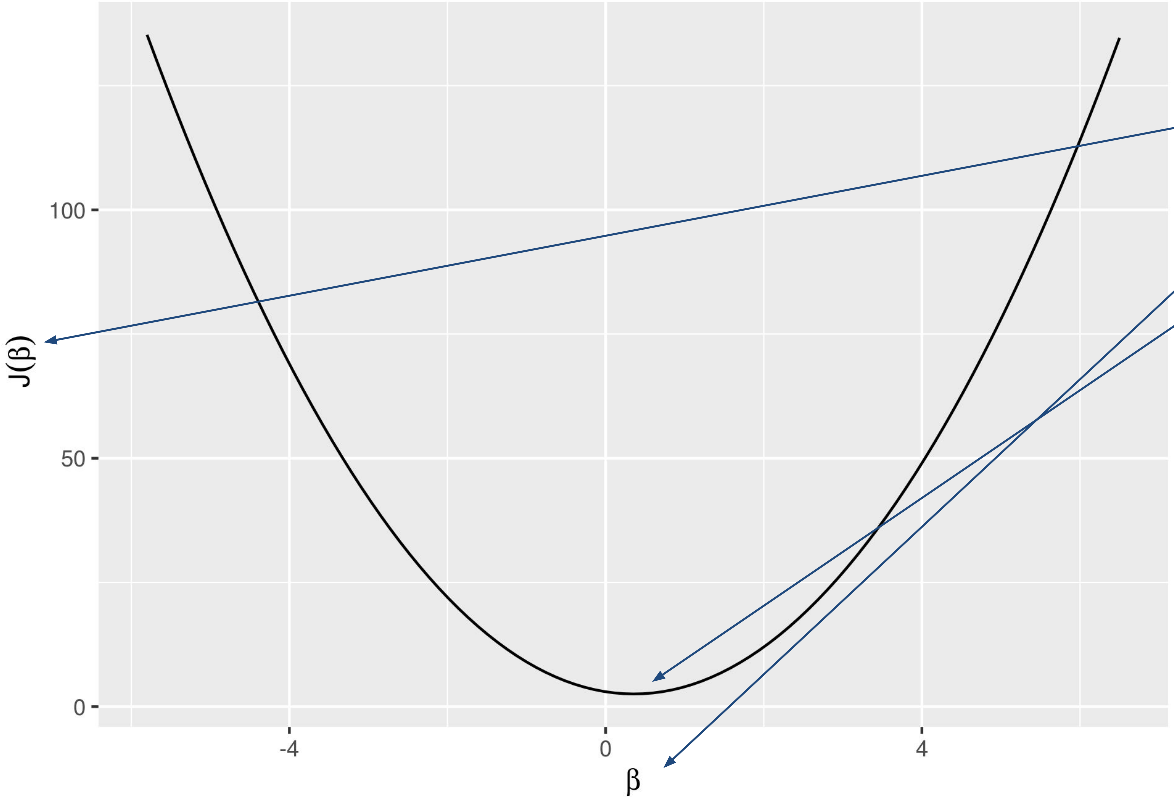
1. Maximise the likelihood function

$$\mathcal{L}(\beta, \sigma^2, y, X) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum(y_i - x_i\beta)^2\right)$$

2. Minimise the loss function

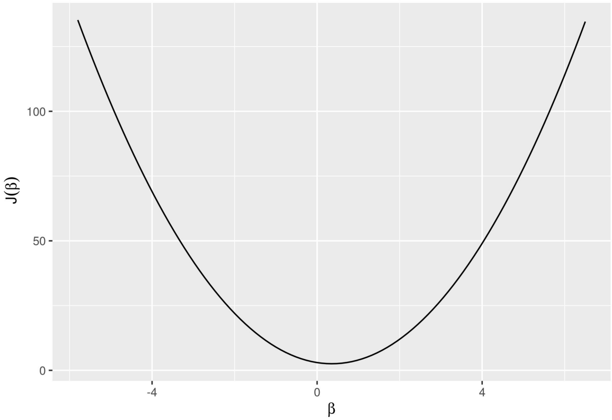
$$J(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(\beta_i X_i - y_i)^2$$

Minimise loss (maximise likelihood)



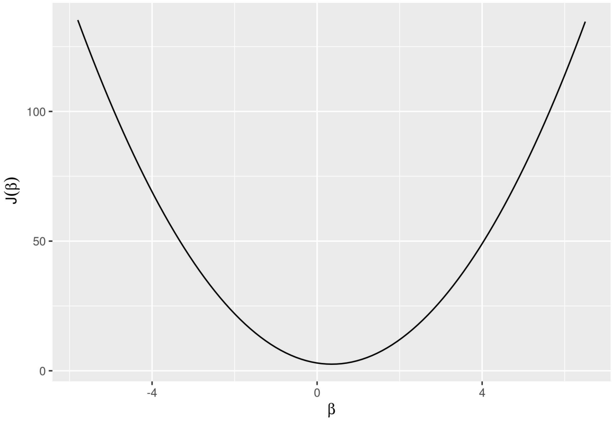
1. loss function
2. model parameters
3. minimum

Minimise loss (maximise likelihood)



1. Start with initial values for β : (initialisation)
2. Change β in the direction of reducing J(β) / increasing L(β) : (descent/ascent)
3. Stop when the minimum/maximum is reached : (min/maximisation)

Minimise loss (maximise likelihood)



This is an iterative process that progressively updates the coefficients (unlike least squares that offers a closed form solution)!

1. Start with initial values for β : (initialisation)
2. Change β in the direction of reducing J(β) / increasing L(β) : (descent/ascent)
3. Stop when the minimum/maximum is reached : (min/maximisation)

How reliable are the estimated coefficients?

- Standard errors of the coefficients

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2 = Var(e)}{\sum(x_i - \bar{x})^2}} = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}$$

- Confidence intervals

$$\hat{\beta}_1 \pm z \cdot \mathrm{SE}(\hat{\beta}_1)$$

- Hypothesis test

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \longrightarrow \text{p-value}$$

Model accuracy

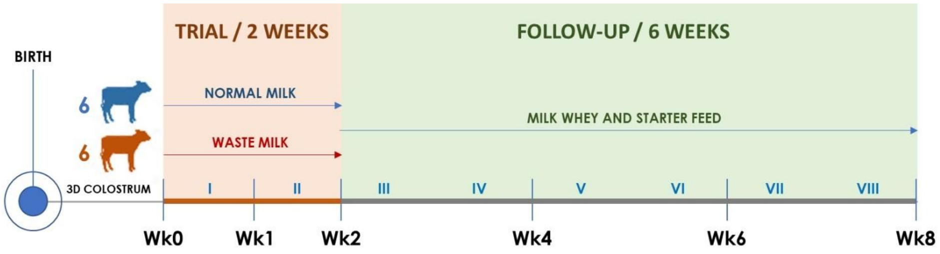
$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$$

Residual standard error: Q: what value do we wish for it?

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Coefficient of determination: normalized RSE Q: what value do we wish for it?

Treatments and timepoints: linear regression



From: Penati et al. 2021 (Front. Vet. Sci., 08 July 2021)

Apply linear regression within-time point:

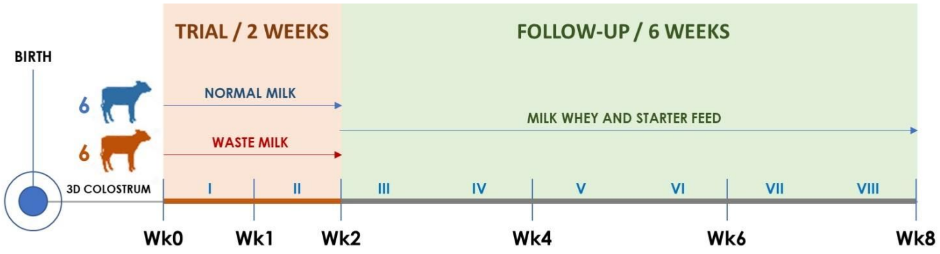
y = b0 + b1*treatment + e

replicate the model for each time point

PROS:
- simple to run
- easy to interpret
- may be your research hypothesis

CONS:
- smaller sample size
- fail to account for changing effect of treatment over time
- may not be your research hypothesis

Across-timepoints analysis



From: [Penati et al. 2021] (Front. Vet. Sci., 08 July 2021)

Previously, we saw that we can fit separate regression models for each timepoint

- each separate regression equation ignores the other timepoints: if timepoints are correlated (time trend) this can lead to misleading estimates of the association between treatment and response
- it can be difficult to combine results and interpret results across timepoints (independent models)

Across-timepoints analysis

$$y = \beta_0 + \beta_1 x_1 + + \beta_2 x_2 + e$$

$$\text{response} = \beta_0 + \beta_1 \text{timepoint} + \beta_2 \text{treatment} + e$$

Across-timepoints analysis

$$\text{response} = \beta_0 + \beta_1 \text{timepoint} + \beta_2 \text{treatment} + e$$

We interpret coefficients "ceteris paribus":

- b0: mean of reference classes (e.g. controls, T0)
- b1: mean(T1) - mean(T0) [and so on]
- b2: mean(treated) - mean(controls)
- b0 + b1: mean(controls) at T1
- b0 + b2: mean(treated) at T0
- b0 + b1 + b2: mean(treated) at T1

Interaction model

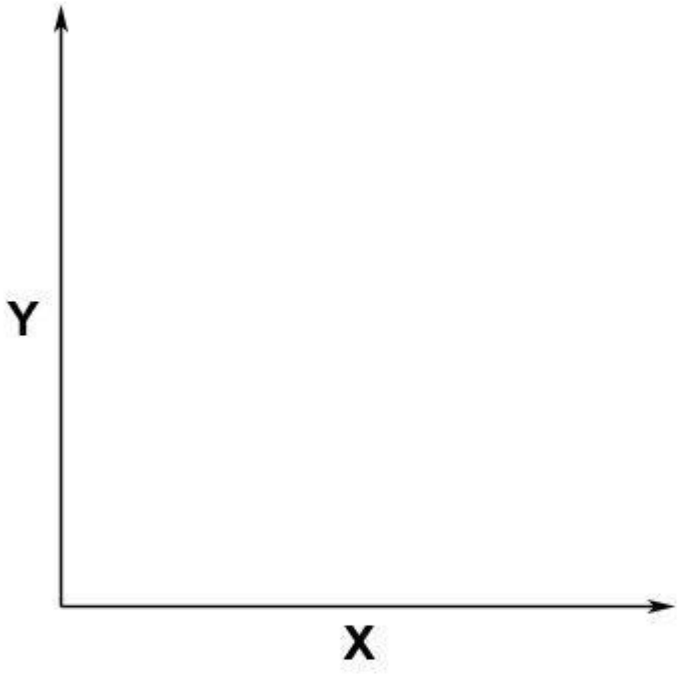
Removing the **additive assumption**: x1 and x2 are assumed to be independent
But are they?

$$y = \beta_0 + \beta_1 x_1 + + \beta_2 x_2 + \beta_3 (x_1 \cdot x_2) + e$$

$$\text{resp} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{treat} + \beta_3 (\text{time x treat}) + e$$

Interaction model

Interaction plot



Interaction model

hierarchical principle : if we include an interaction in a model we always have to also include the main effects, even if the p-values associated with their coefficients are not significant

(no models with only interaction terms!)