Lecture 4: Regression

Bob O'Hara

bob.ohara@ntnu.no

Regression

Now we will get to modelling

Two Reasons for Models

Inference

- Does giving a hurricane a male name increase the amount of damage it does?
- How much difference is there between hurricanes with male or female names?

Prediction

- if we call the next hurricane Donald, how much damage will it do?

The Model Idea

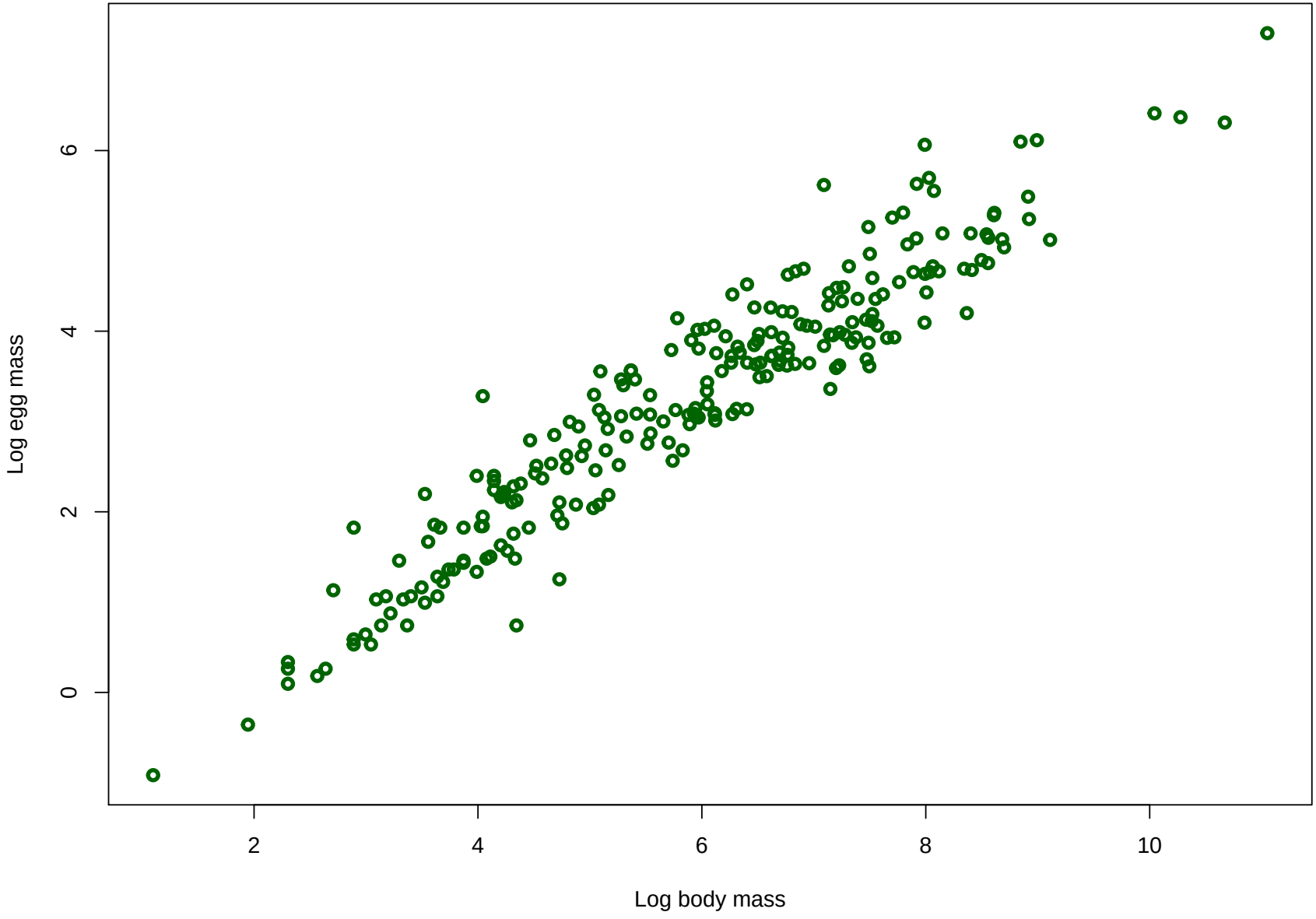
We have a variable, Y, which we want to explain with a covariate, X

e.g. we want to explain egg size in species of bird by body size

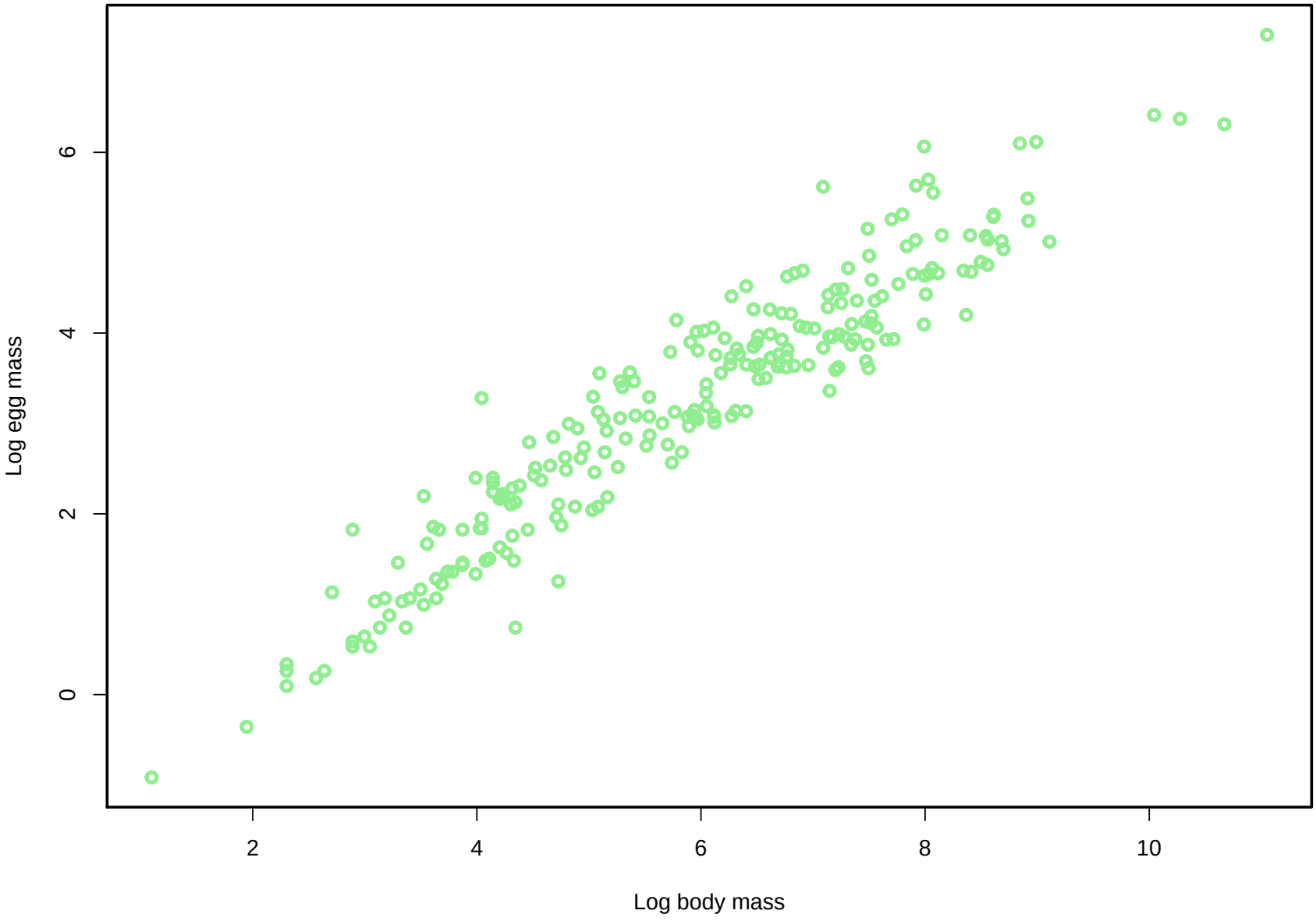
Here we will use a straight line

- more complex models come later: they are sums of straight lines

Some Data

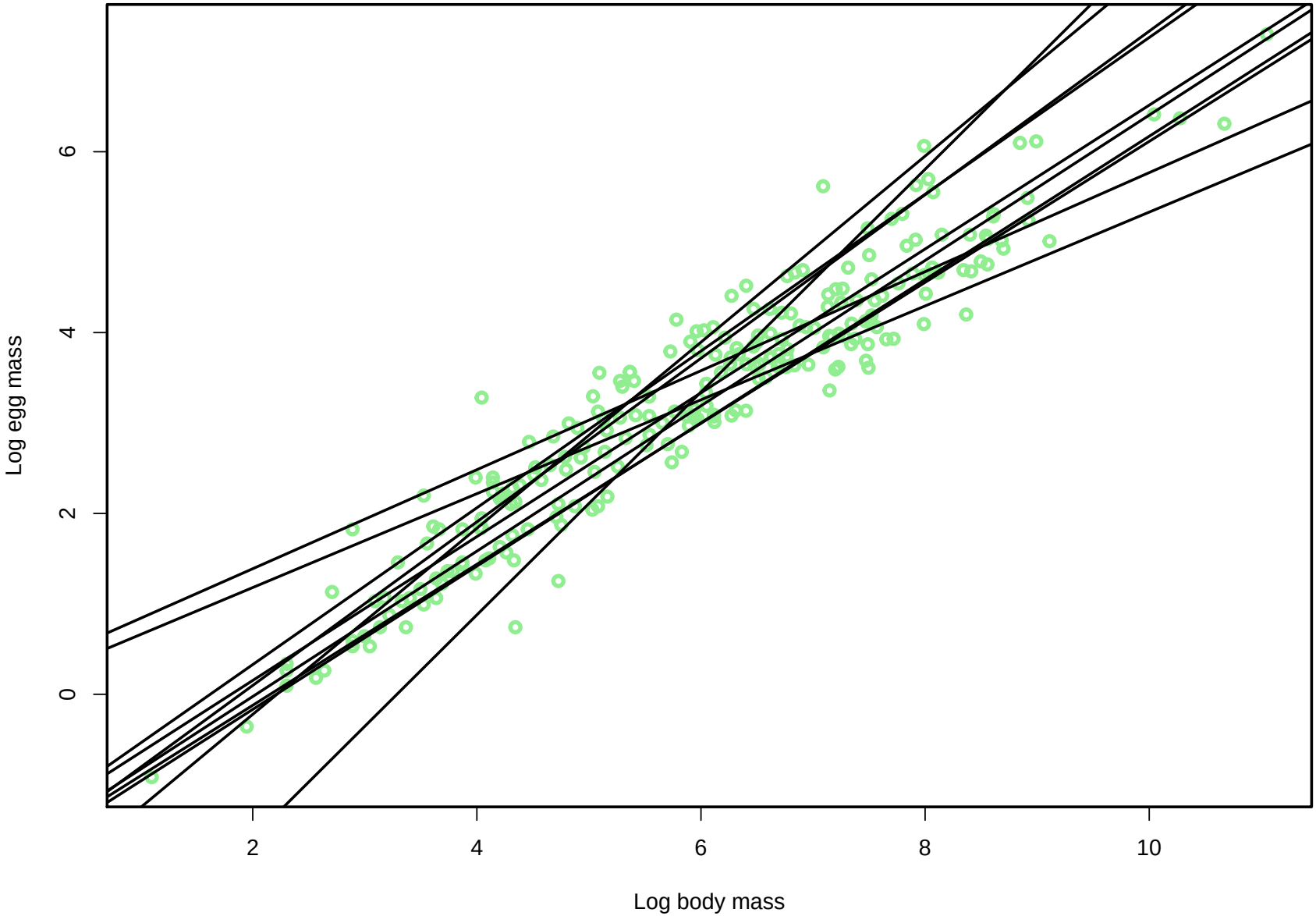


What is the best line?



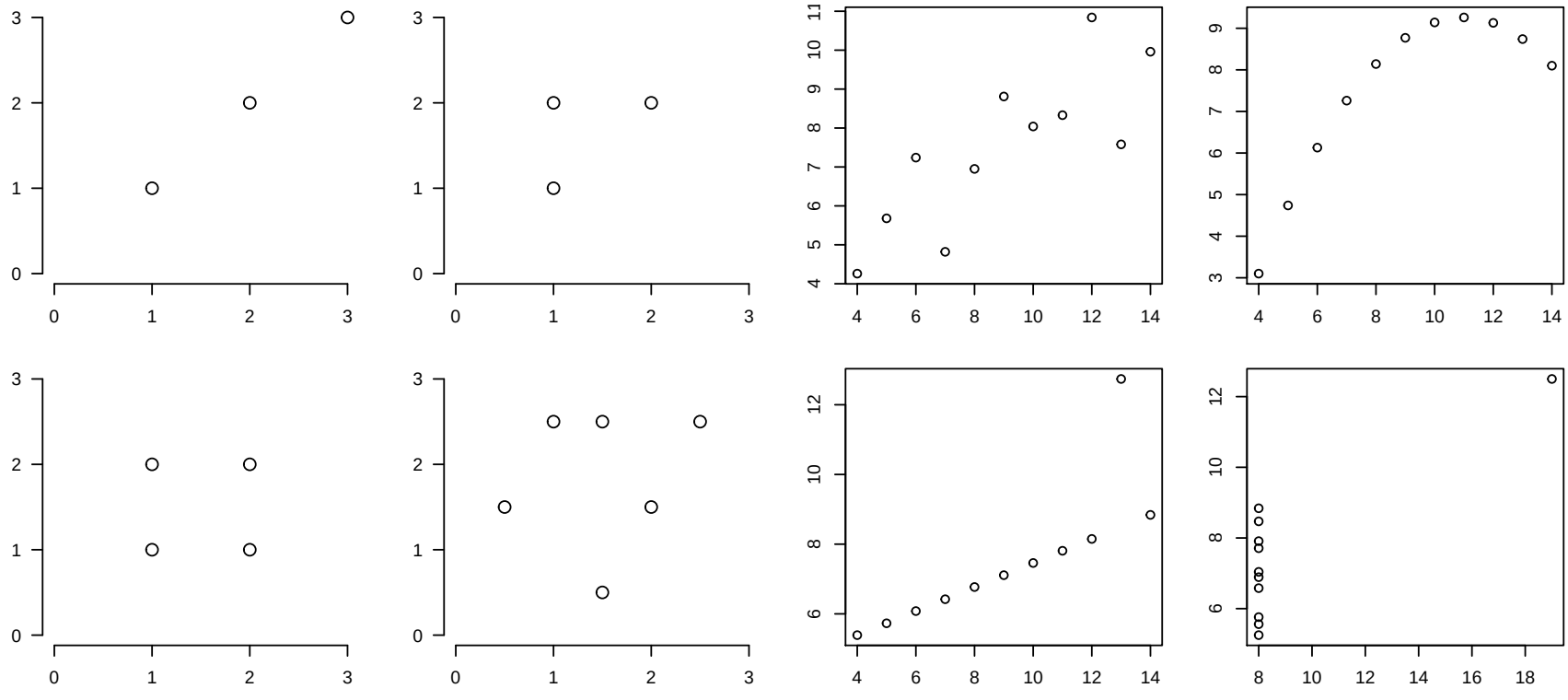
What is the best line?

One of these?



Excercise: What should a straight line be?

For each plot, draw what you think is the best line



Defining a best line

We want to fit a straight line:

$$y_i = \alpha + \beta x_i$$

But there is error that the line cannot explain, so we change the model to

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

with $\sum \varepsilon_i = 0$

Maximum Likelihood

We can assume that the errors are normally distributed:

$$\varepsilon_i \sim N(0, \sigma^2)$$

The log-likelihood is

$$l(\mathbf{y}|\mathbf{x}, \alpha, \beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}$$

Maximum Likelihood

We can assume that the errors are normally distributed:

$$\varepsilon_i \sim N(0, \sigma^2)$$

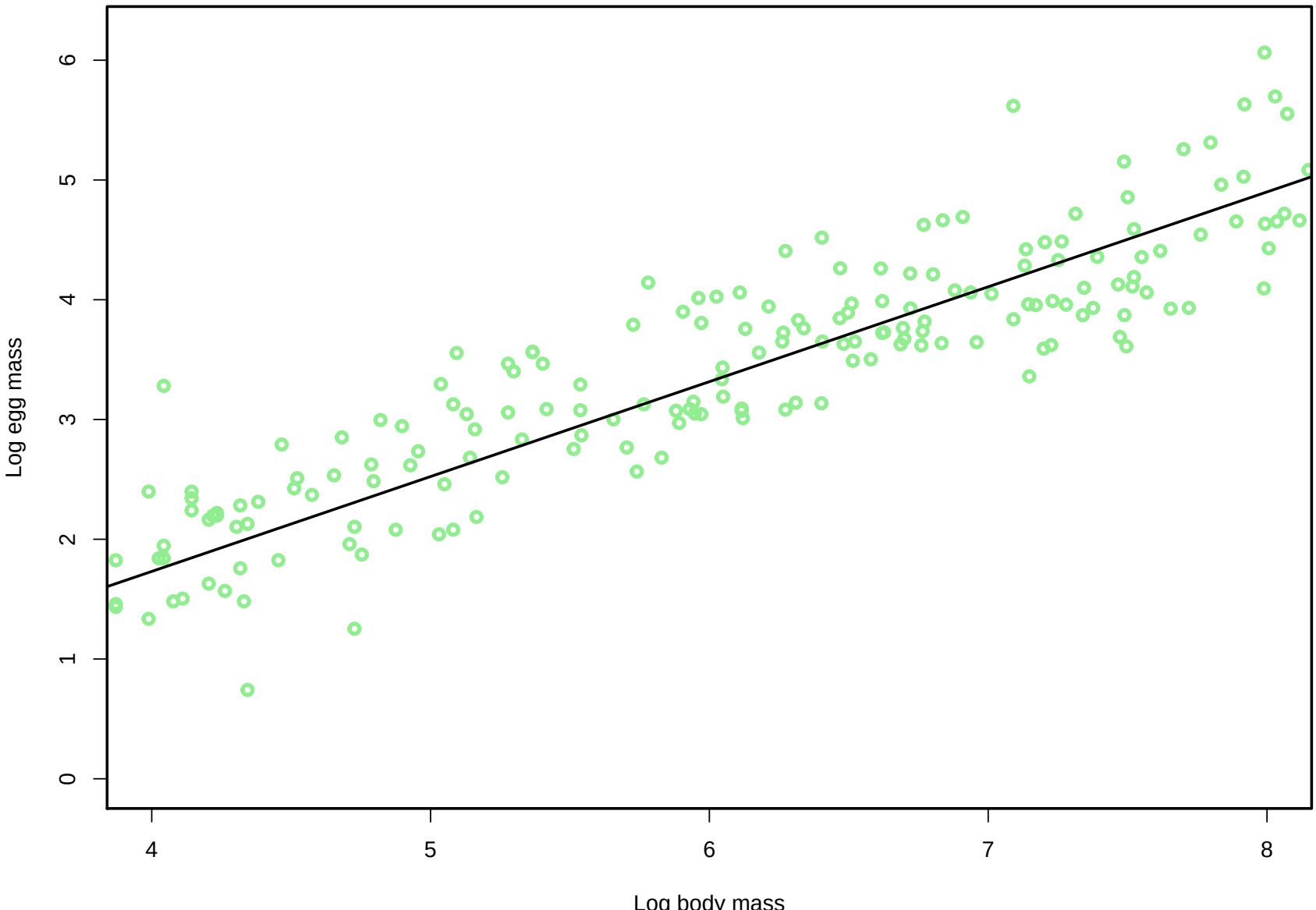
The log-likelihood is

$$l(\mathbf{y}|\mathbf{x}, \alpha, \beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}$$

This is quadratic in y_i, so this is the same minimising the sums of squares, i.e. the least squares estimate

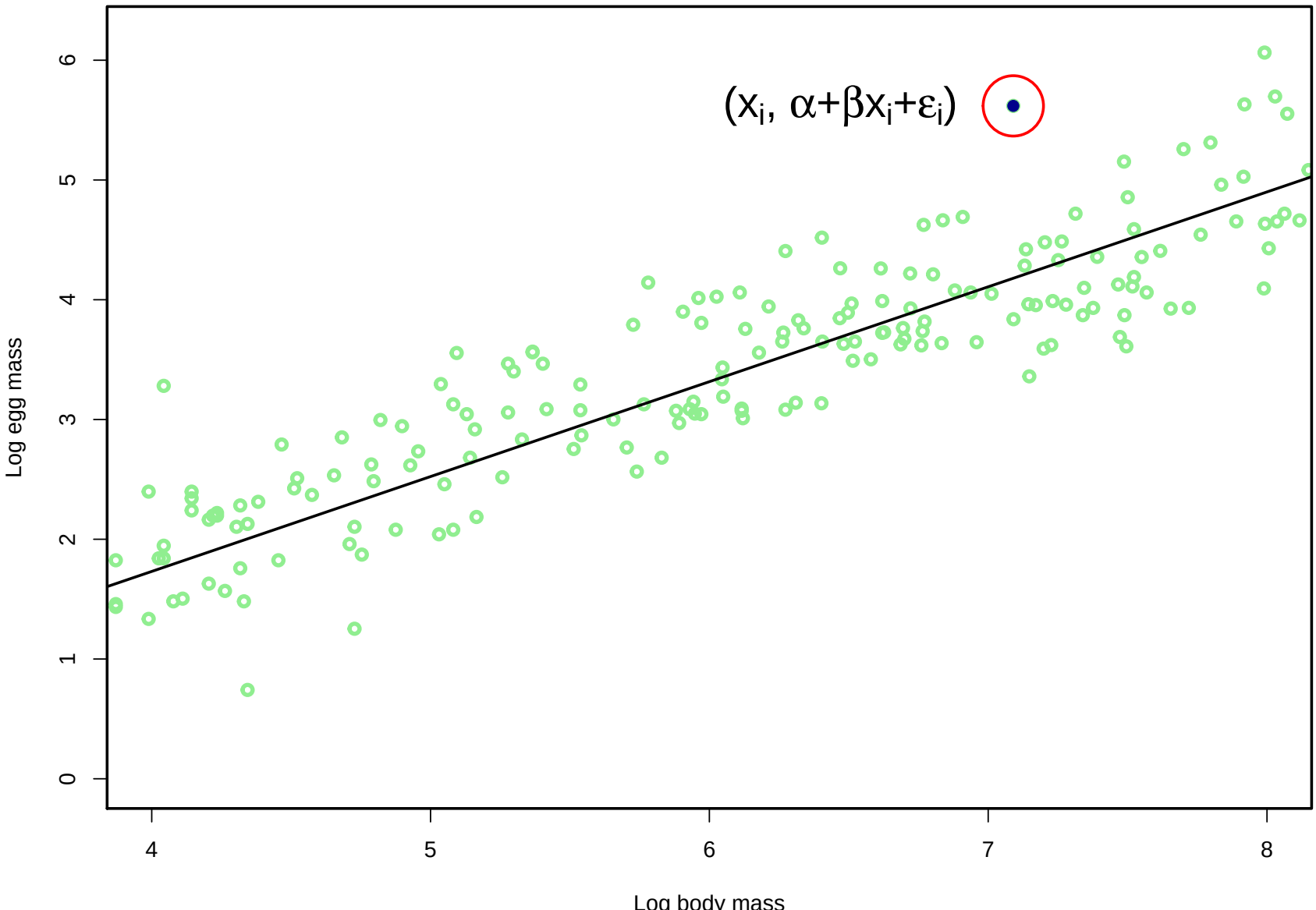
Drawing a best line

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



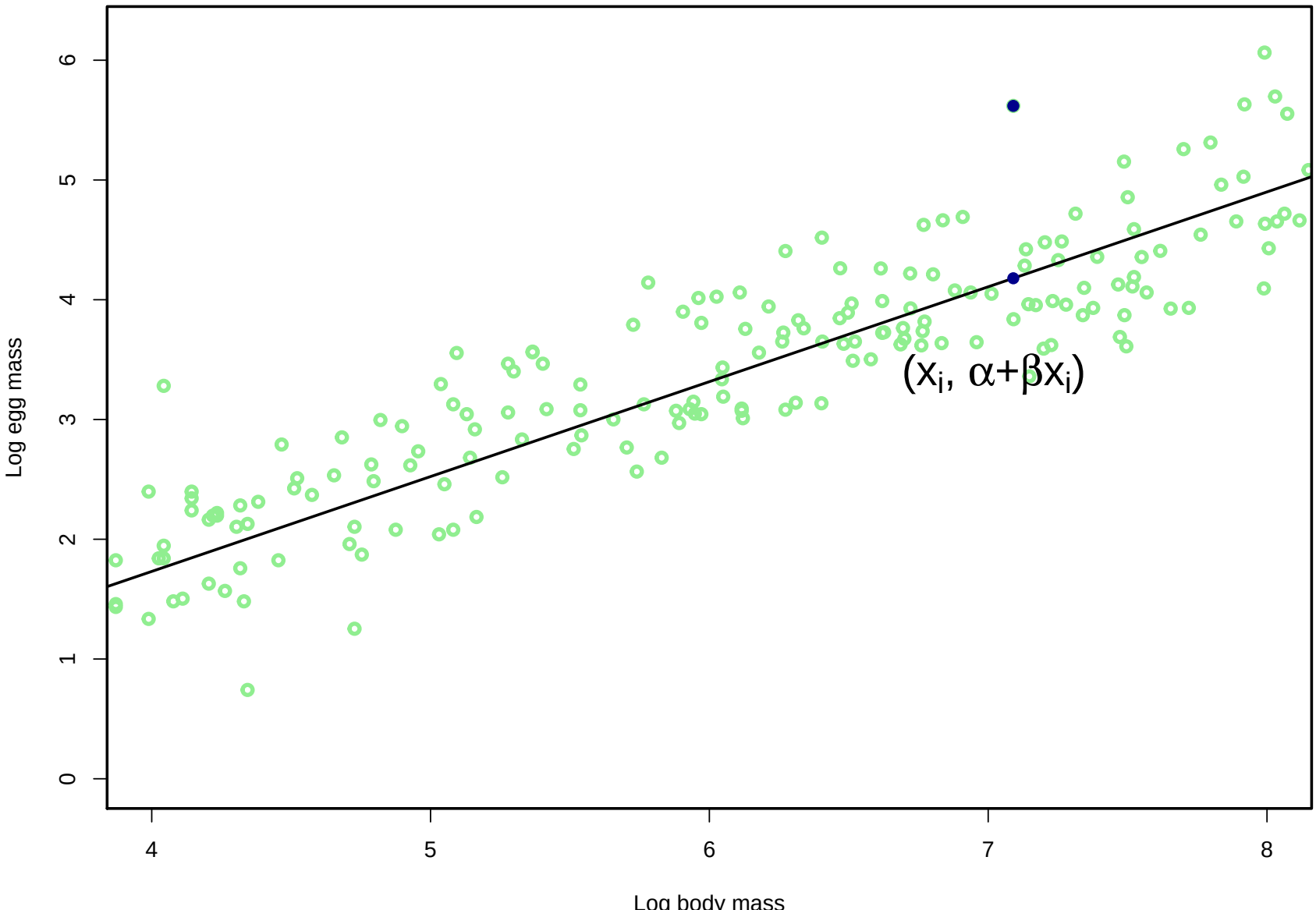
Drawing a best line

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



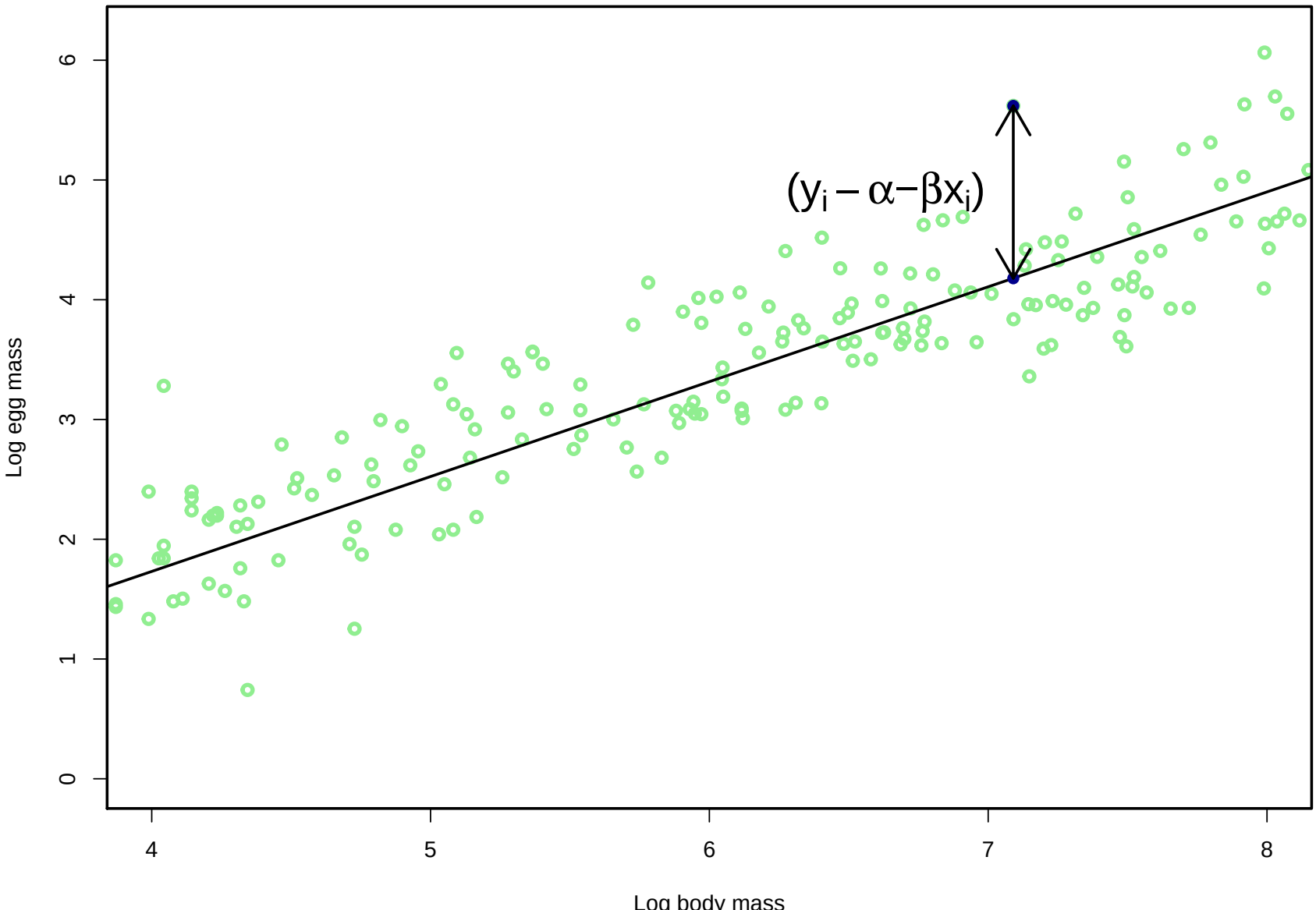
Drawing a best line

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



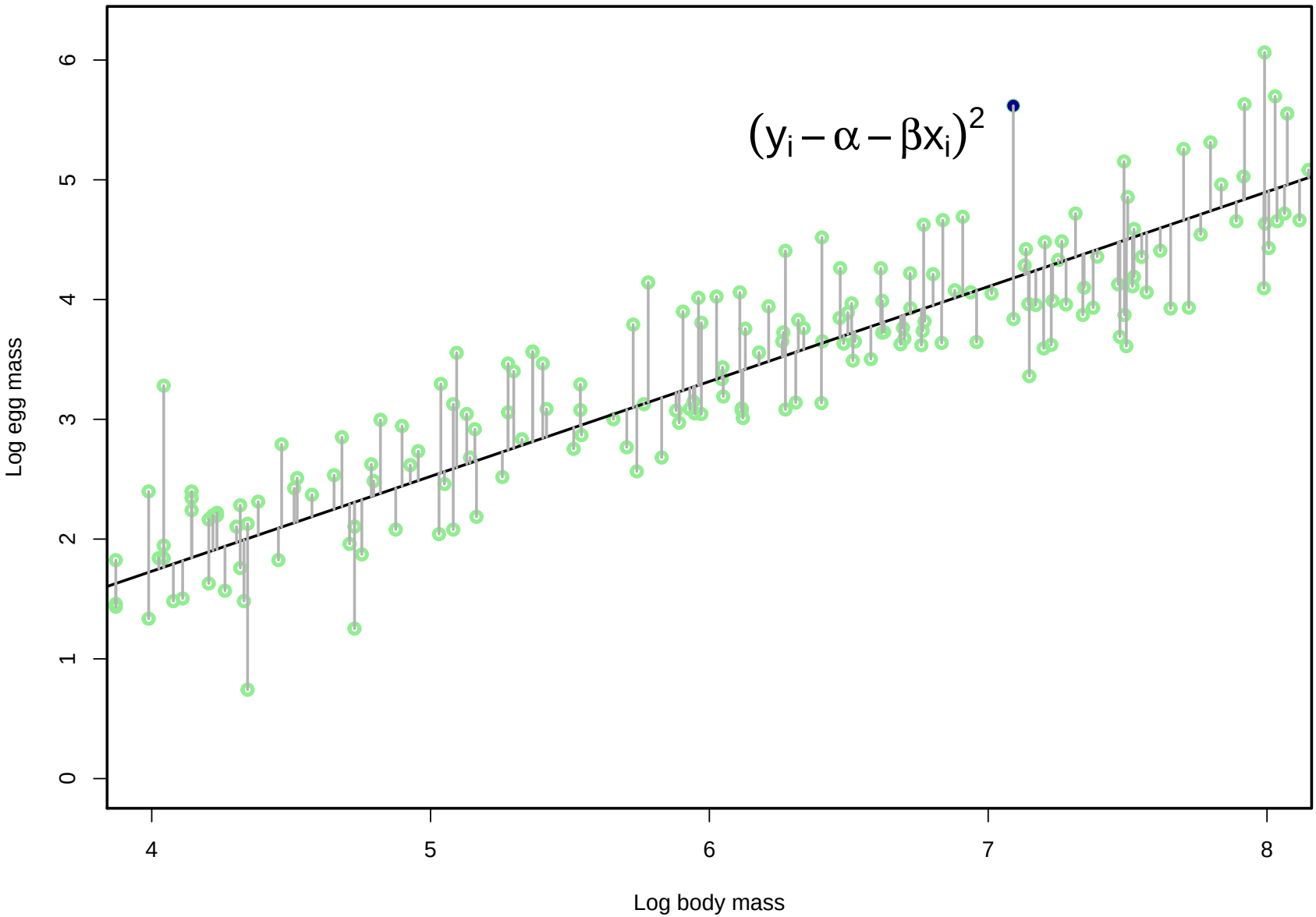
Drawing a best line

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



Drawing a best line

We minimise the squares of these distances



MLEs

We could go through the maths, but

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

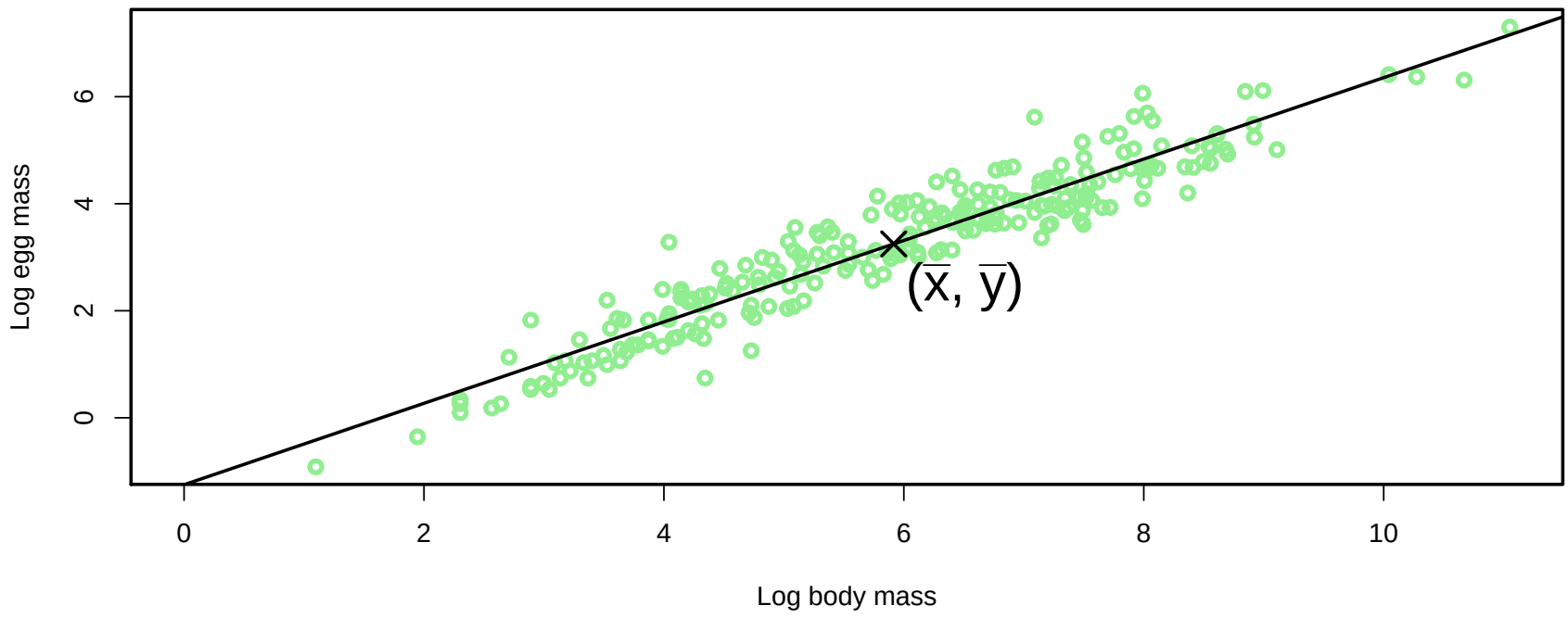
$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$$

What these Mean: The Intercept

The line goes through $(\hat{x}, \hat{y})$, and is extrapolated backwards to $x = 0$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$



What these Mean: The slope

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

This is $Cov(x, y)/Var(x)$

What these Mean: the variance

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$$

Set $\hat{\mu_i} = \hat{\alpha} + \hat{\beta}x_i$ and we get

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu_i})^2$$

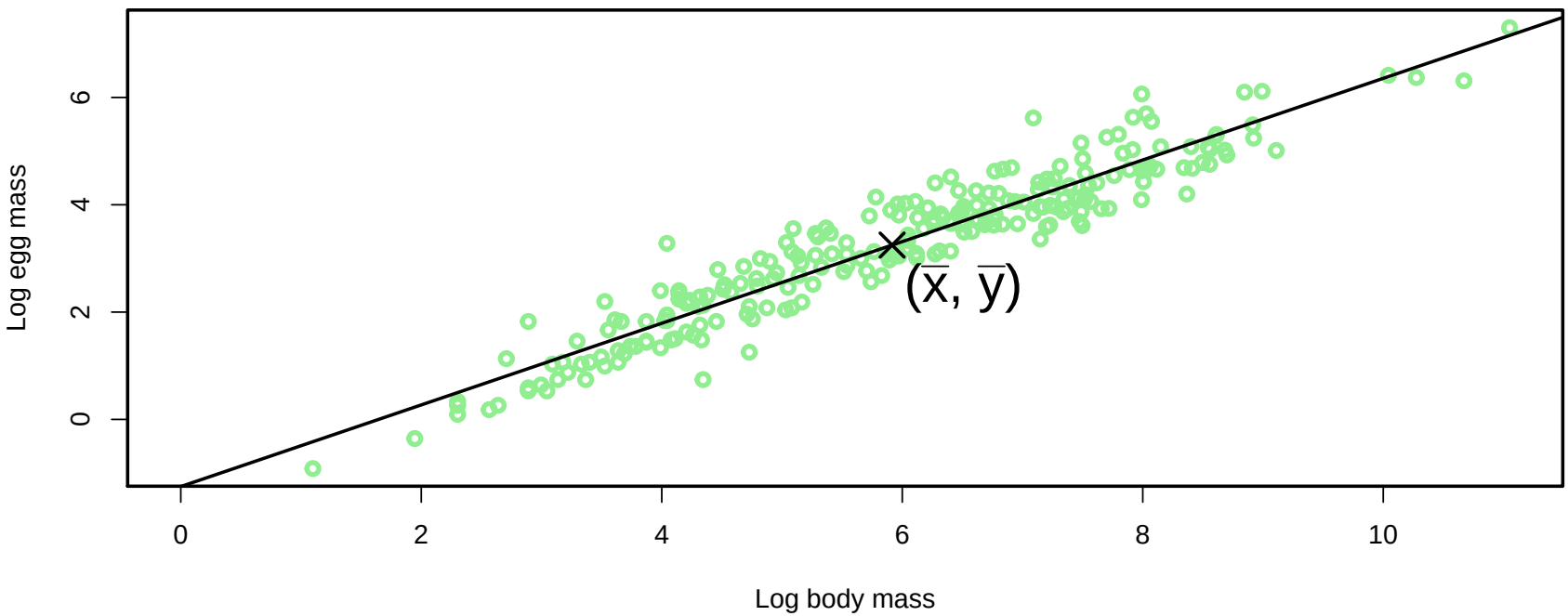
So, the same as the ML estimate of the variance, once we have corrected the values to their means

The Example

For this data we have

$y_i = \text{-}1.25 + 0.76\ x_i + \varepsilon_i$

$\varepsilon_i \sim N(0,\ \ 0.46^{\,2})$



Interpreting The Example

For this data we have

y_i = -1.25 + 0.76 $x_i + \varepsilon_i$

$\varepsilon_i \sim N(0,\ 0.46^{\,2})$

If log Body Mass increases by 1

- body mass increases by $e^1 = 2.72$ times
- egg mass increases by 0.76 (i.e. 2.14 times)

so this is a bit less than proportional: it is about 3/4

Mis-Interpreting The Example

$y_i = \text{-}1.25 + 0.76\ x_i + \varepsilon_i$

$\varepsilon_i \sim N(0,\ \ 0.46^{\,2})$

A mass-less bird would have an egg of negative mass

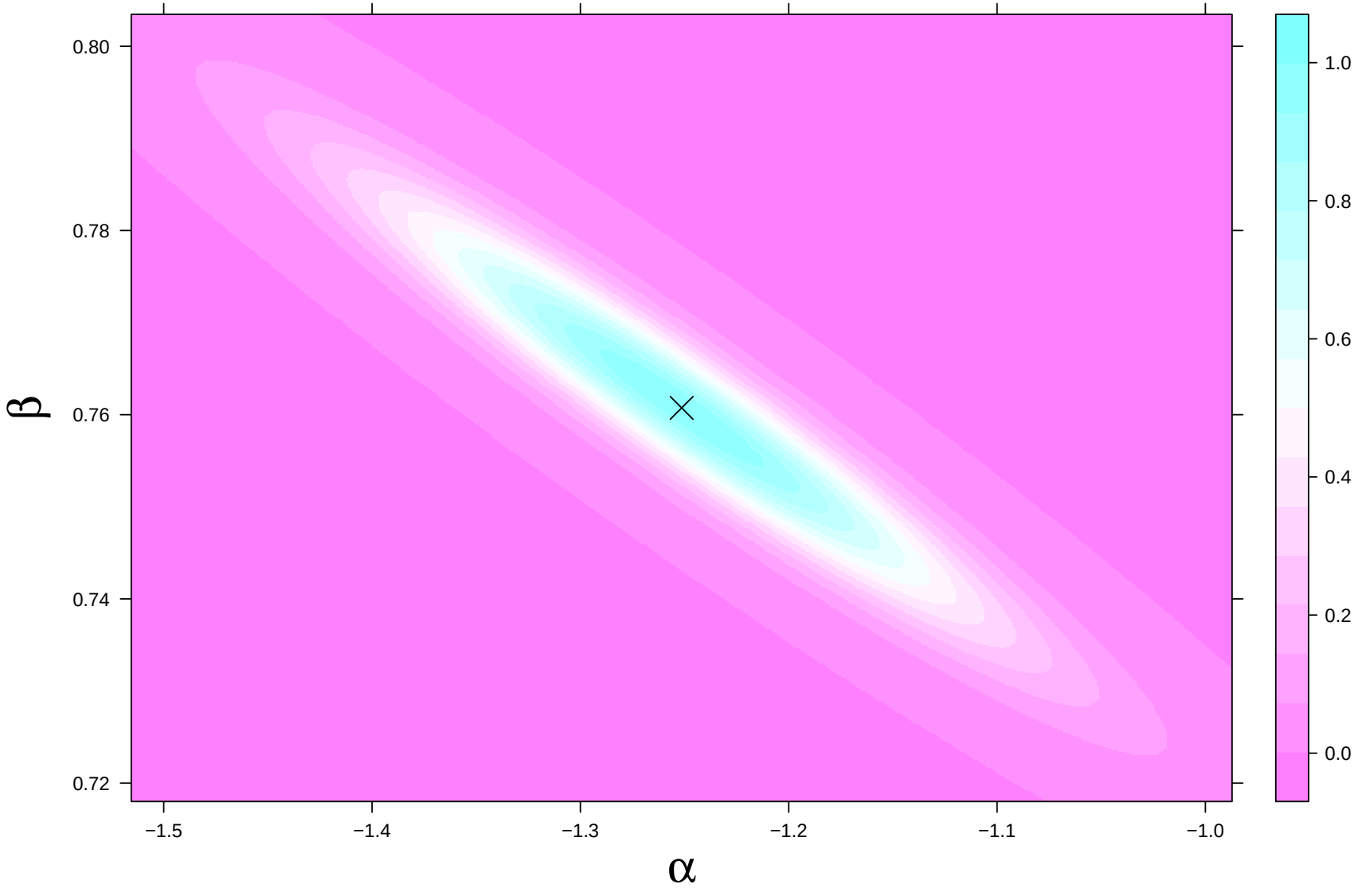
- why does the model still make sense?

How good is the model?

- how good (i.e. precise) are the estimates?
- how well does it explain the data?

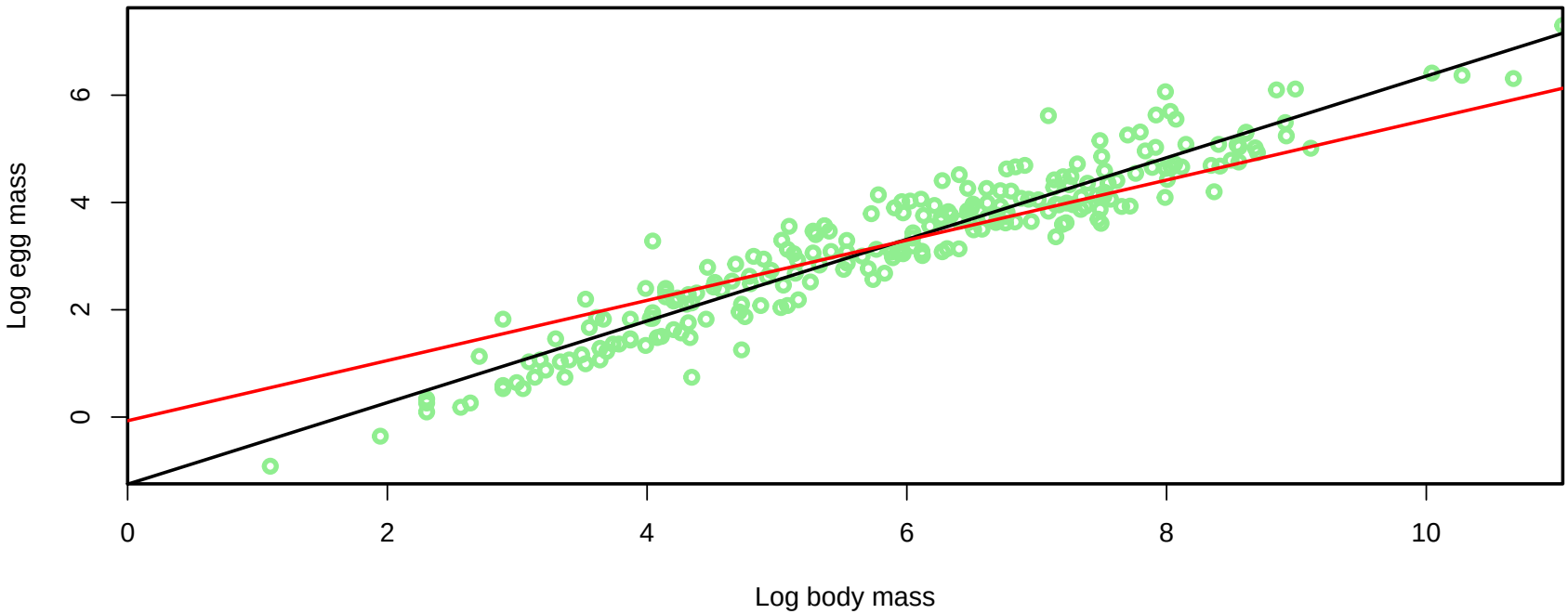
How precise are the estimates?

The likelihood for α and β is correlated



Why is there a correlation?

The line has to go through $(\bar{x}, \bar{y})$, so if we increase the intercept, we have to decrease the slope



Standard Errors

The standard errors are not trivial to obtain, especially as they are correlated. We are primarily interested in the standard errors for α and β

$$Var\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

Note that the covariance is 0 when $\sum x_i = 0$

Confidence Intervals

From the standard errors we can calculate the asymptotic confidence intervals for the parameters, e.g.

$$(\hat{\alpha} - 1.96 s_\alpha, \hat{\alpha} + 1.96 s_\alpha)$$

where s_α is the standard error for α (in practice, get this from the computer)

e.g the 95% confidence interval for the slope is (0.73, 0.79), i.e. we can be sure it is < 1, and is fairly close to $3/4$

(with less data, a t-distirbution makes more sense)

Prediction

Our Eclectus, Freyja and Eric, are thinking of breeding (at least Freyja is). How large will their egg be?

Prediction

Female Eclectus' body size: 390g - 445g (so median of 417.5g)

Point estimate:

$E(y) = -1.25 + 0.76 \times log(417.5) = 3.34$

So, a mass of exp(3.34) = 28.2g

But this is uncertain...

(remember, we have log-log transformed the data)

Prediction Intervals

Just as we have confidence intervals, we can also have prediction intervals

We assume our data is normally distributed (given its mean), so if we knew the prameters it would be

$$y_{pred} \sim N(\alpha + \beta x_{pred}, \sigma^2)$$

e.g. the 95% confidence interval would be
$e^{(3.34-1.96\times 0.46, 3.34+1.96\times 0.46)} = (11.5, 68.8)$

Full Prediction Intervals

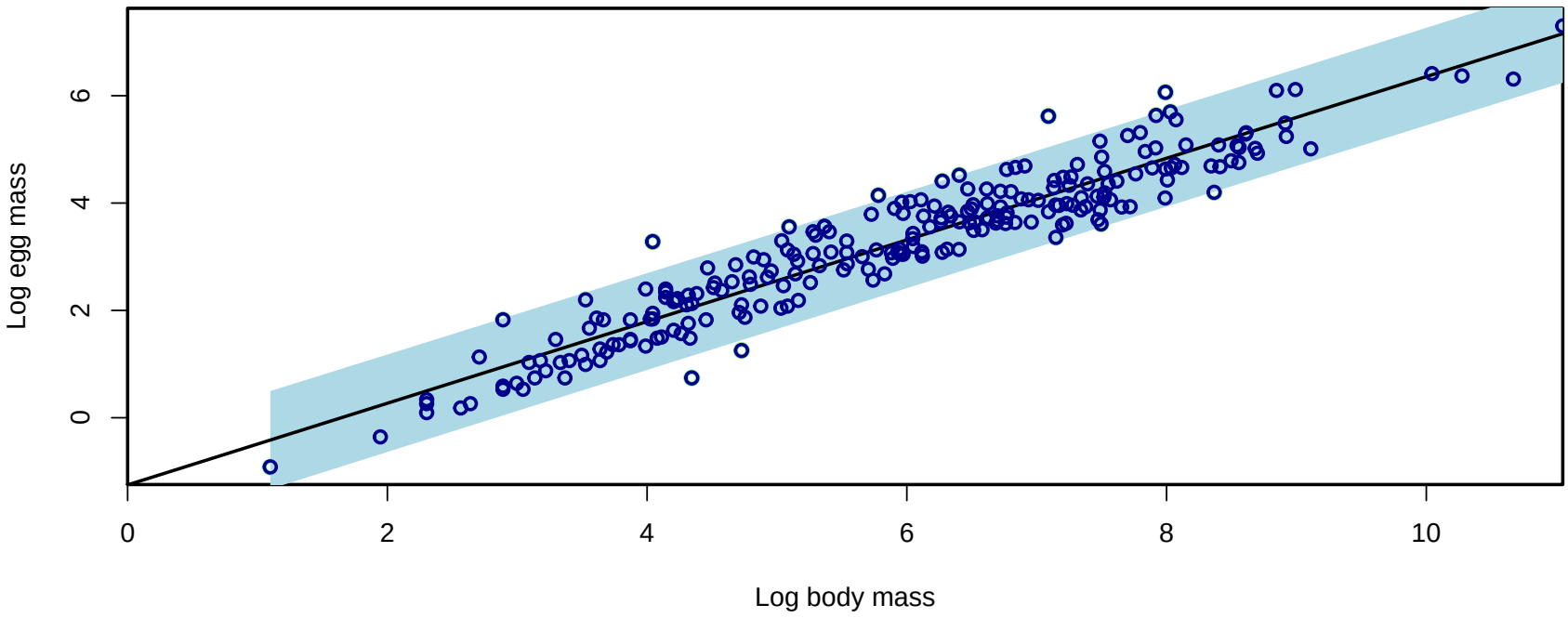
We also take into account the uncertainty in the data (and thus parameter estimates), so the standard deviation of the prediction becomes

$$\sigma_{pred} = \sqrt{1 + \frac{1}{n} + \frac{(x_{pred} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

We also have to assume a t-ditribution (with n-1 degrees of freedom), to account for the uncertainty in $\hat{\sigma}^2$

$$y_{pred} \sim t(\alpha + \beta x_{pred}, \sigma^2, n - 1)$$

Full Prediction Intervals



Next...

Friday: Exercises (at last!)

`https://www.math.ntnu.no/emner/TMA4268/2018v/1Intro/Rbeginner.html`

for masochists: `https://www.math.ntnu.no/emner/TMA4268/2018v/1Intro/Rintermediate.html`

Next Week: How well does the model actually fit the data?

- Grafen & Hails Chapter 2.4 - 2.7