Stat 412/512

LINEAR COMBINATIONS AND PREDICTIONS

Jan 21 2015

Charlotte Wickham

stat512.cwick.co.nz

Two big types of question

Find the effect of .... e.g. the x + 1 type

Find the mean when .... e.g. the x= 0 type

Just applications of those

Find the difference in mean between
....

Which parameter captures ...

Case Study 10.2 Echolocation

Some bats use echolocation to orient themselves.

Echolocation is energy expensive but maybe some bats have evolved to do it efficiently.

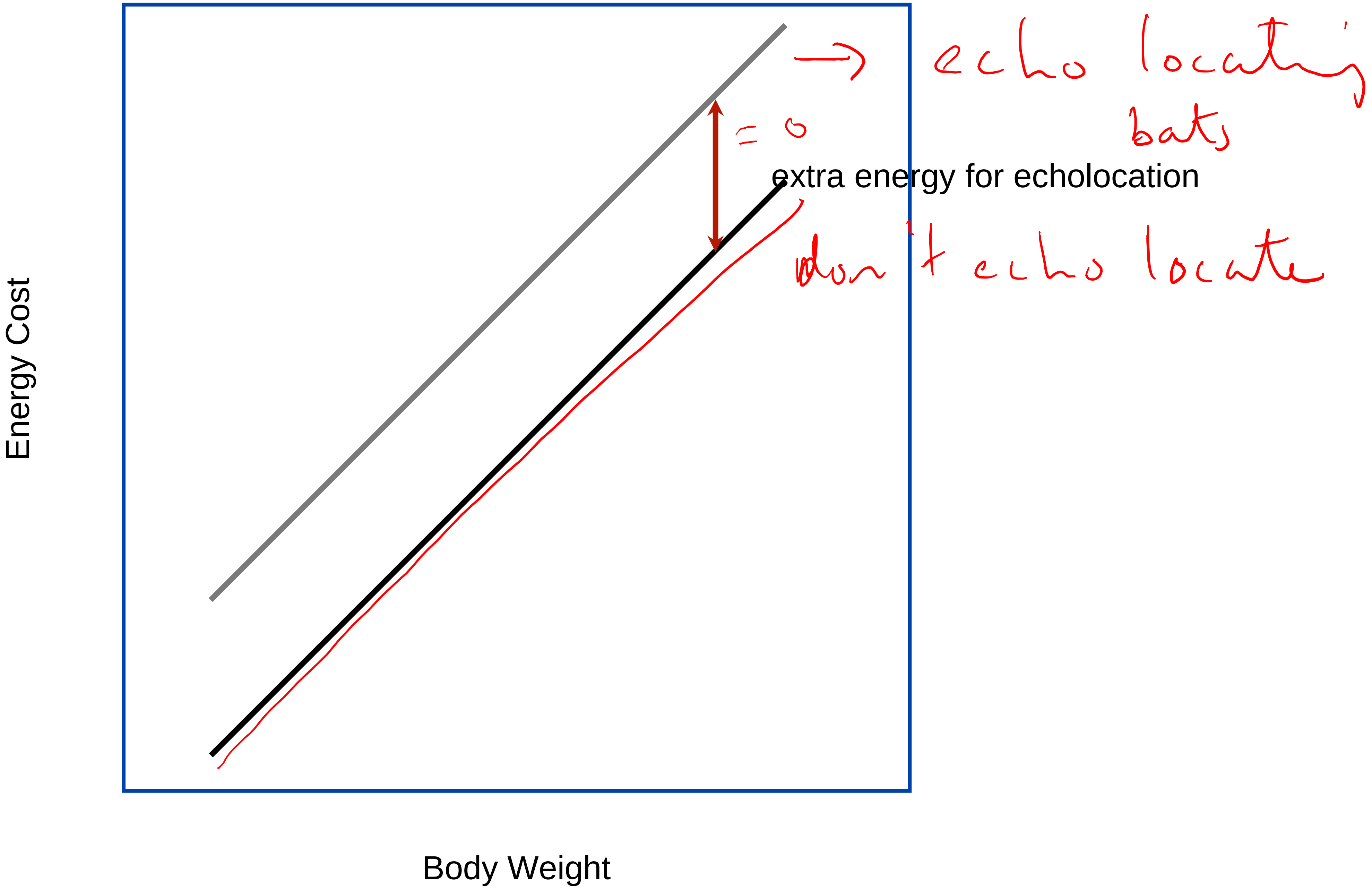
Zoologists wonder whether the energy costs of echolocation during flight are the sum of flights costs plus echolocation.

Cost during flight = cost of flight + cost of stationary echolocation

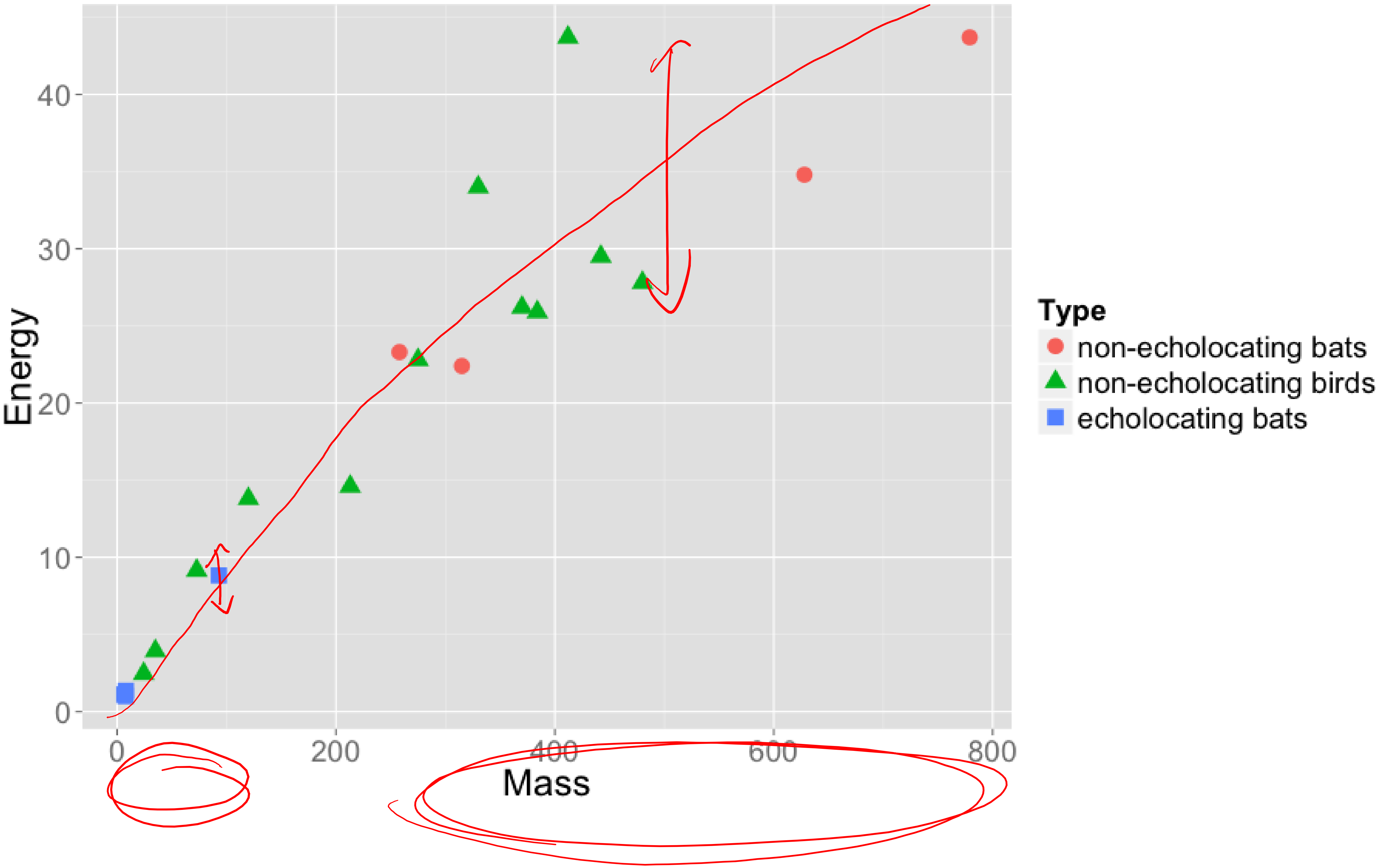
Complication: the energy costs of flight depend on how heavy you are

Heavy bats expend more energy flying.

But, for bats of the same body weight, echolocating bats should expend a constant amount of energy more than non-echolocating bats.

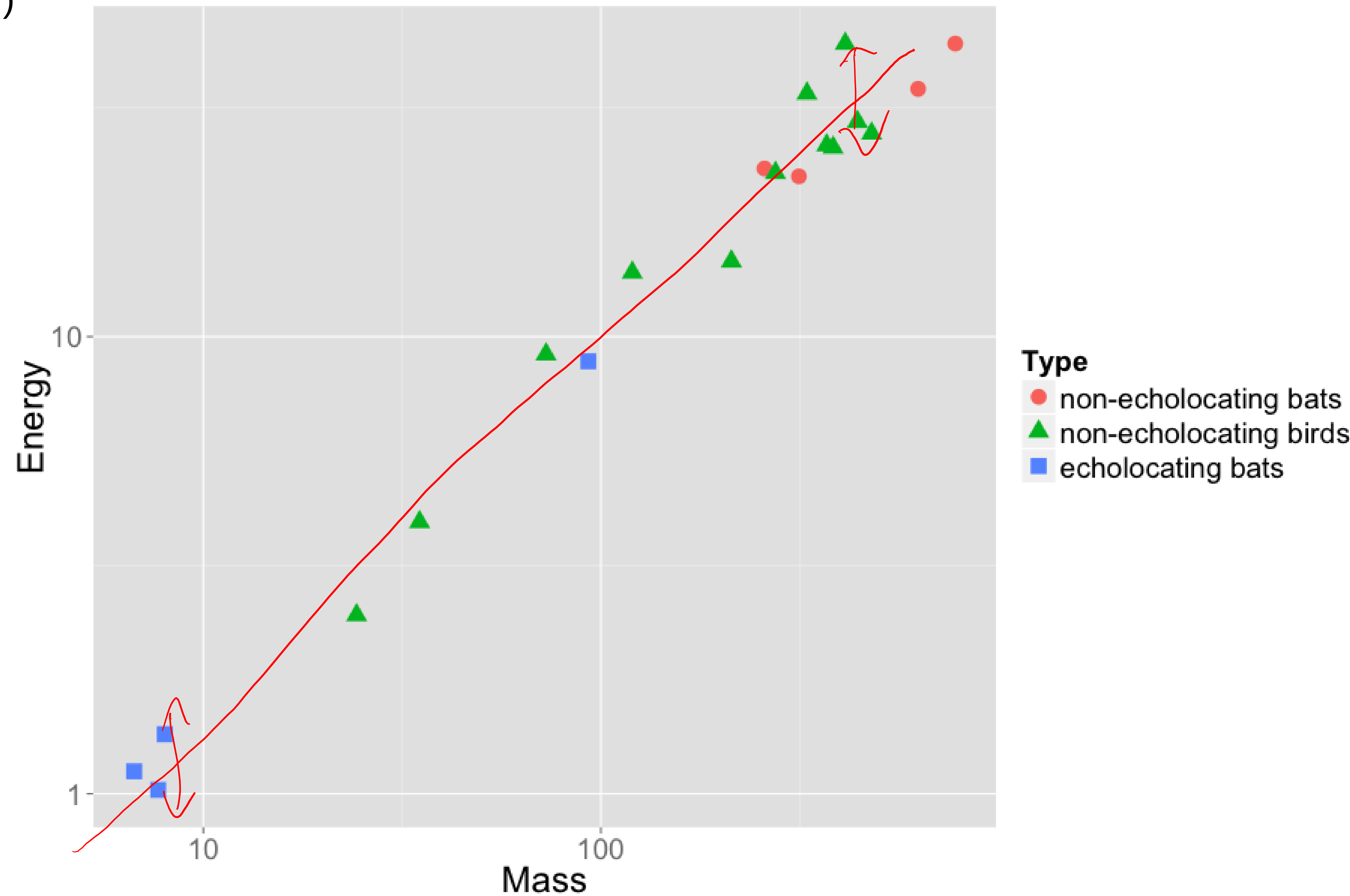


qplot(Mass, Energy, data = case1002, colour = Type, shape = Type)



Mass and inflight energy from 20 energy studies

birds help to define cost to weight relationship qplot(Mass, Energy, data = case1002, colour = Type, shape = Type, log = "xy")



log transformed: removes curvature and non-constant variation

A tentative model

$\mu\{$ log Energy | log Mass, Type$\}$

$= \quad$ log Mass + TYPE shorthand

$= \quad \beta_0 + \beta_1 \log \text{Mass} + \beta_2\, bird + \beta_3\, ebat$

where,

ebat is an indicator for echolocating bat,

bird is an indicator for bird

The easiest way to understand a model with indicator variables in it, is to write out the model within each category,

for non-echolocating bats

$\mu\{\log \text{Energy} \mid \log \text{Mass}, \text{ebat} = 0, \text{bird} = 0\} =$

$= \beta_0 + \beta_1 \log \text{Mass}$

intercept slope

for echolocating bats

$\mu\{\log \text{Energy} \mid \log \text{Mass}, \text{ebat} = 1, \text{bird} = 0\} =$

$= (\beta_0 + \beta_3) + \beta_1 \log \text{Mass}$

for birds:

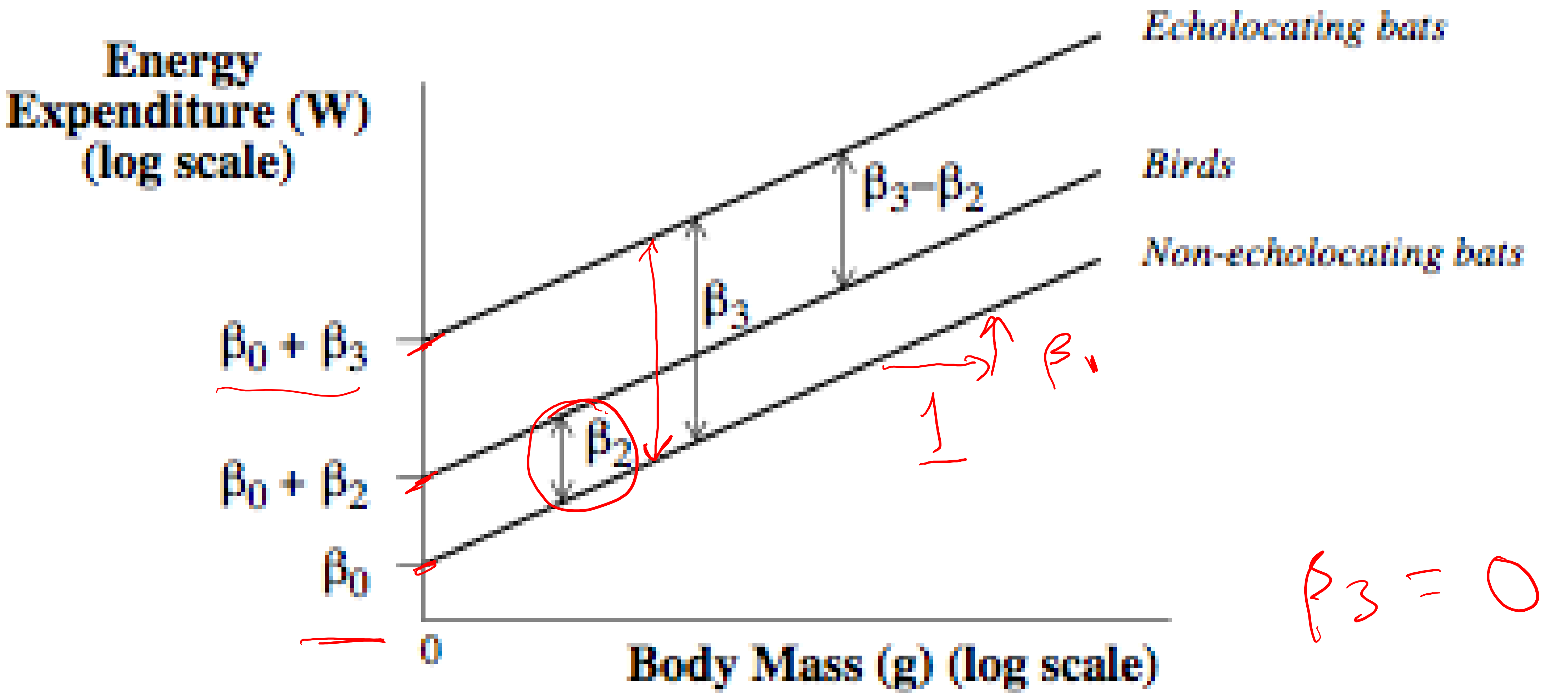
$\mu\{\log \text{Energy} \mid \log \text{Mass}, \text{ebat} = 0, \text{bird} = 1\} =$

$= (\beta_0 + \beta_2) + \beta_1 \log \text{Mass}$

A parallel lines model with three categories

Display 10.5 **p. 272**

The parallel regression lines model for the bat echolocation data



Does the model answer the question of interest?

Yes,

if $\beta_3 > 0$ echolocation while flying is associated with an extra β_3 in mean log energy.

if $\beta_3 = 0$ echolocation while flying is not associated with any extra mean log energy.
(The bats have evolved to be efficient).

We can answer our question of interest with a test with the null, $\beta_3 = 0$.

Is the model appropriate for our data?

You might ask whether a separate lines model is more appropriate.

$\mu\{\text{log Energy} \mid \text{log Mass, Type}\}$

$$= \text{log Mass} + \text{TYPE} + \text{log Mass} \times \text{TYPE}$$

$$= \beta_0 + \beta_1 \text{ log Mass} + \beta_2\, bird + \beta_3\, ebat +$$

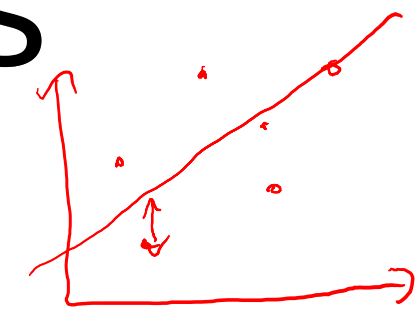
$$\cancel{\beta_4\, ebat \times log\ Mass} + \cancel{\beta_5\, bird \times log\ Mass}$$

We could test the null hypothesis $\beta_4 = \beta_5 = 0$, the relationship between body mass and energy costs doesn't depend on type

Inference on more than one parameter, next week

You should also ask if the assumptions of multiple linear regression are appropriate (Chapter 11).

Estimation of parameters



Just like in simple linear regression, the parameters are estimated by minimizing the sum of the squared residuals, a.k.a **least squares**

The formulas for the estimates are best represented using matrix algebra (see ex 10.20 & 10.21).

estimate

Notation: $\hat{\beta}_j$ is the least squares estimate of β_j, the j'th coefficient in the model.

true value unknown

Estimate of σ

We assume constant spread about the regression line, σ and estimate σ, with

$$\hat{\sigma} = \sqrt{\frac{\text{Sum of squared residuals}}{\text{Degrees of freedom}}}$$

Degrees of freedom = n - # of β

sample size

number of parameters

In ecolocation study: n = 20, parallel lines model has 4 β's,

$\beta_0 + \beta_1 \log \text{Mass} + \beta_2\, ebat + \beta_3\, bird$

d.f. = 20 - 4 = 16

Fact

Assuming the response is Normally distributed with constant spread, σ, at each combination of the explanatory variables,

$$\text{t-ratio} = \frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}_{\hat{\beta}_j}}$$

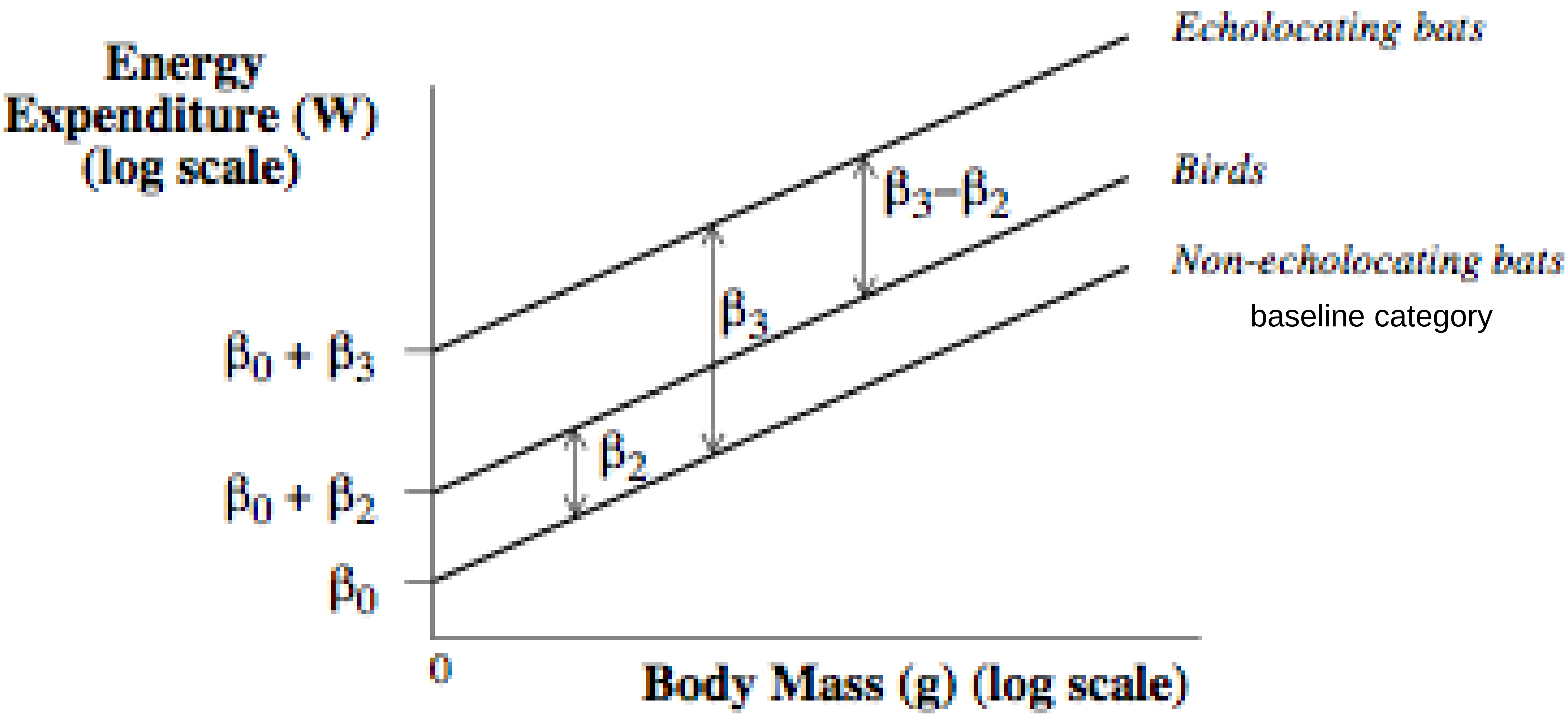
has a **Student's *t*-distribution** with degrees of freedom equal to the degrees of freedom associated with $\hat{\sigma}$**.**

, $n - \#$ params

There are formulas for SE($\hat{\beta}_i$), the standard error of our estimate.

Display 10.5 **p. 272**

The parallel regression lines model for the bat echolocation data



$\mu\{\log \text{Energy} \mid \log \text{Mass}, \text{Type}\} = \beta_0 + \beta_1 \log \text{Mass} + \beta_2\, bird + \beta_3\, ebat$

$\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}_{\hat{\beta}_j}}$ has a Student's t-distribution

Leads to tests and confidence intervals

$$\frac{\hat{\beta}_j - 0}{\mathrm{SE}_{\hat{\beta}_j}}$$

To test the null $\underline{\beta_j = 0}$, compare to a Student's t-distribution with d.f. degrees of freedom.

95 % confidence interval for β_j,

$$\hat{\beta}_j \pm \overset{t}{\cancel{q}}_{d.f.}(0.975)\mathrm{SE}_{\hat{\beta}_j}$$

estimate ± multiplier × SE of estimate

d.f. = n - number of β's

Is $\beta_3 = 0$?

```
>  fit_bat <- lm(log(Energy) ~ log(Mass) + Type, data = case1002)
>  summary(fit_bat)

...
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.57636    0.28724  -5.488 4.96e-05 ***
log(Mass)                    0.81496    0.04454  18.297 3.76e-12 ***
Typenon-echolocating birds   0.10226    0.11418   0.896    0.384
Typeecholocating bats        0.07866    0.20268   0.388    0.703
```

Null: $\beta_j = 0$

$$(0.07866 - 0) / 0.20268 = 0.388$$

$$2 * (1 - pt(abs(0.388), 16)) = 0.703$$

There is no evidence that β_3 is not zero.

There is no evidence that echolocating bats expend more energy, after accounting for body mass, than non-echolocating bats (p-value = 0.70).

What is β_3?

0.07866 - qt(0.975, 16) * 0.20268 = -0.3510024

0.07866 + qt(0.975, 16) * 0.20268 = 0.5083224

95% CI for β_3 is -0.35, 0.51

exp(-0.35) = 0.70, exp(0.51) = 1.66

Log transformed Energy & Mass

With 95% confidence the median energy expended by a echolocating bat is between .70 and 1.66 times the median energy expended by non-echolocating bats in this study.

Or: confint(fit_bat)

$-0.35 \qquad 0.51$

With 95% confidence, the mean log Energy for echo-locating bats is between 0.35 units lower & 0.51 units higher than the mean log Energy for non-echolocating bats, after accounting for log Mass.

Significance depends on what is in the model

Three models:

1. μ{ log Energy | log Mass, Type} = TYPE
2. μ{ log Energy | log Mass, Type} = TYPE + log Mass
3. μ{ log Energy | log Mass, Type} = TYPE + log Mass + TYPE x log Mass

1. $= \beta_0 + \beta_1 \, ebat + \beta_2 \, bird$

2. $= \beta_0 + \beta_1 \log Mass + \beta_2 \, bird + \beta_3 \, ebat$

3. $= \beta_0 + \beta_1 \log Mass + \beta_2 \, bird + \beta_3 \, ebat + \beta_4 \, bird \times \log Mass + \beta_5 \, ebat \times \log Mass$

Significance depends on what is in the model

Estimated coefficient on ebat:

1. -2.74, p-value = 0.000259
2. 0.079, p-value = 0.703
3. -1.27, p-value = 0.3406

Interpretation of coefficient on ebat:

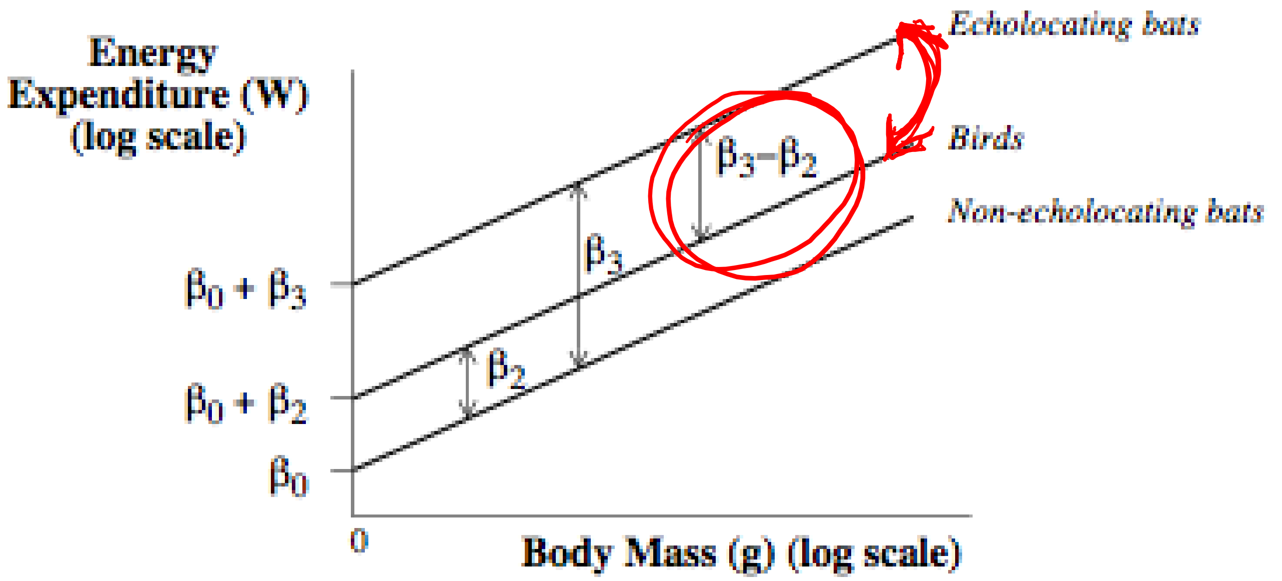
1. difference between mean log energy of ebats and non-ebats ignoring body mass
2. difference between mean log energy of ebats and non-ebats accounting for body mass
3. slopes are different so intercept has quite different meaning, you really need two parameters to characterise the difference between echo-locating bats and non-echolocating bats.

What about the difference between birds and echolocating bats?

Is $\beta_3 - \beta_2 = 0$?

Display 10.5 p. 272

The parallel regression lines model for the bat echolocation data



Either:
work out standard error on difference between coefficients 10.4.3 not examinable

or redefine reference level

In the parallel lines (and separate lines) models, the parameters are relative to the reference category.

Redefine model

* just to indicate the βs in this model might not be the same as our other model

μ{ log Energy | log Mass, Type}

$$= \text{log Mass} + \text{TYPE}$$

$$= \beta_0^* + \beta_1^* \text{ log Mass} + \beta_2^* \textit{ non-ebat} + \beta_3^* \textit{ bird}$$

The parallel regression lines model for the bat echolocation data