Stat 412/512

Starting Inference

Jan 16 2015

Charlotte Wickham

stat512.cwick.co.nz

Announcements

Quiz #1 Friday (Jan 23rd) next week in class. No notes, no book, you wont need a calculator.

Practice questions posted on website

You can expect about 3 three questions of that length.

Case Study 9.2 Mammalian Brain Size

Big brains are better, but come with costs.

We know bigger animals would have bigger brains in general, but if we could remove that effect, what else would be related to larger brains?

Observed average brain weight, body weight, gestation length and litter size for 96 mammals.

What characteristics are associated with large brains, after accounting for body size?

Average values of brain weight, body weight, gestation length, and litter size in 96 species of mammal

Species	Brain Weight (grams)	Body Weight (kilograms)	Gestation Period (days)	Litter Size
Quokka	17.5	3.5	26	1.0
Hedgehog	3.50	0.93	34	4.6
Tree shrew	3.15	0.15	46	3.0
Elephant shrew I	1.14	0.049	51	1.5
Elephant shrew II	1.37	0.064	46	1.5
Lemur	22.	2.1	135	1.0
Slow loris	12.8	1.2	90	1.2
Bush baby	9.9	0.7	135	1.0
Howler monkey	54.	7.7	139	1.0
Ring-tail monkey	73.	3.7	180	1.0
Spider monkey I	114.	9.1	140	1.0
Spider monkey II	109.	7.7	140	1.0
Gentle lemur	7.8	0.22	145	2.0
Rhesus monkey I	84.6	6.0	175	1.0
Rhesus monkey II	107.	8.7	165	1.1
Hamadryas baboon	183.	21.	180	1.0
Western baboon	179.	32.	180	1.0
Vervet guenon	67.	4.6	195	1.0
Leaf monkey	65.5	5.8	168	1.0
White-handed gibbon	102.	5.5	210	1.0
Orangutan	343.	37.	270	1.0
Chimpanzee	360.	45.	230	1.0
Gorilla	406.	140.	265	1.0
Human being	1,300.	65.	270	1.0
Long-nosed armadillo	12.	3.7	120	4.0
Aardvark	9.6	2.2	31	5.0
Jack rabbit	13.3	2.9	41	2.5
Tree squirrel	6.23	0.33	38	3.0
Flying squirrel	1.89	0.052	40	3.1
Canadian beaver	40.	20.	128	2.9
Beaver	45.	25.	128	4.0
Deer mouse I	0.68	0.027	23	3.7
Deer mouse II	0.63	0.026	23	5.0
Deer mouse III	0.52	0.017	24	5.0
Deer mouse IV	0.69	0.024	24	5.0
Hamster I	0.67	0.036	21	4.6
Hamster II	1.12	0.13	16	6.3
Pygmy gerbil	1.04	0.065	21	4.0
Rat I	0.72	0.05	23	7.3
Rat II	2.38	0.34	21	8.0
House mouse	0.45	0.024	19	5.0
Hopping mouse	1.18	0.15	27	5.6
Porcupine I	37.	11.	112	1.2
Porcupine II	37.	14.	112	1.2
Porcupine III	24.	6.6	113	1.0
Guinea pig	4.28	0.97	67	2.6
Capybara	76.	30.	123	3.0
Agoutis	20.3	2.8	104	1.3

Species	Brain Weight (grams)	Body Weight (kilograms)	Gestation Period (days)	Litter Size
Acouchis	9.9	0.78	98	1.2
Chinchilla	5.25	0.43	110	2.0
Nutria	23.	5.0	132	5.5
Dolphin	1,600.	160.	360	1.0
Porpoise	537.	56.	270	1.0
Dog	70.2	8.5	63	4.0
Red fox	48.	6.0	52	4.0
Gray fox	37.3	3.8	63	3.7
Bat-eared fox	28.5	3.2	65	4.0
Grizzly bear	400.	250.	219	2.3
Beaked whale	500.	250.	240	1.8
Raccoon	41.6	5.3	63	3.5
Kinkajou	[illegible]	[illegible]	[illegible]	[illegible]
Badger	[illegible]	[illegible]	[illegible]	[illegible]
Domestic ca	[illegible]	[illegible]	[illegible]	[illegible]
Lynx	[illegible]	[illegible]	[illegible]	[illegible]
Leopard	[illegible]	[illegible]	[illegible]	[illegible]
Lion	[illegible]	[illegible]	[illegible]	[illegible]
Tiger	[illegible]	[illegible]	[illegible]	[illegible]
Fur seal	[illegible]	[illegible]	[illegible]	[illegible]
Sea lion	[illegible]	[illegible]	[illegible]	[illegible]
Harp seal	[illegible]	[illegible]	[illegible]	[illegible]
Weddell sea	[illegible]	[illegible]	[illegible]	[illegible]
African Ele	[illegible]	[illegible]	[illegible]	[illegible]
Hyrax	[illegible]	[illegible]	[illegible]	[illegible]
Horse	[illegible]	[illegible]	[illegible]	[illegible]
Tapir	[illegible]	[illegible]	[illegible]	[illegible]
Wild boar	[illegible]	[illegible]	[illegible]	[illegible]
Domestic pig	180.	190.	115	8.0
Hippopotamus	590.	1,400.	240	1.0
Pygmy hippopotamus	260.	150.	205	1.0
Llama	225.	93.	330	1.0
Vicuna	198.	45.	300	1.1
Barking deer	124.	16.	183	1.1
Fallow deer	223.	80.	240	1.0
Axis deer	219.	89.	218	1.0
Red deer	435.	200.	255	1.0
Elk	365.	120.	235	1.0
Sambar	383.	120.	246	1.1
Caribou	288.	110.	225	1.0
Eland	480.	560.	255	1.0
Yak	334.	250.	255	1.0
Cattle	456.	520.	280	1.0
Duikers	93.	13.	120	1.0
Blackbuck Antelope	200.	39.	180	1.0
Barbary sheep	210.	66.	158	1.2
Domestic sheep	125.	49.	150	2.4
Domestic goat	106.	30.	151	2.0

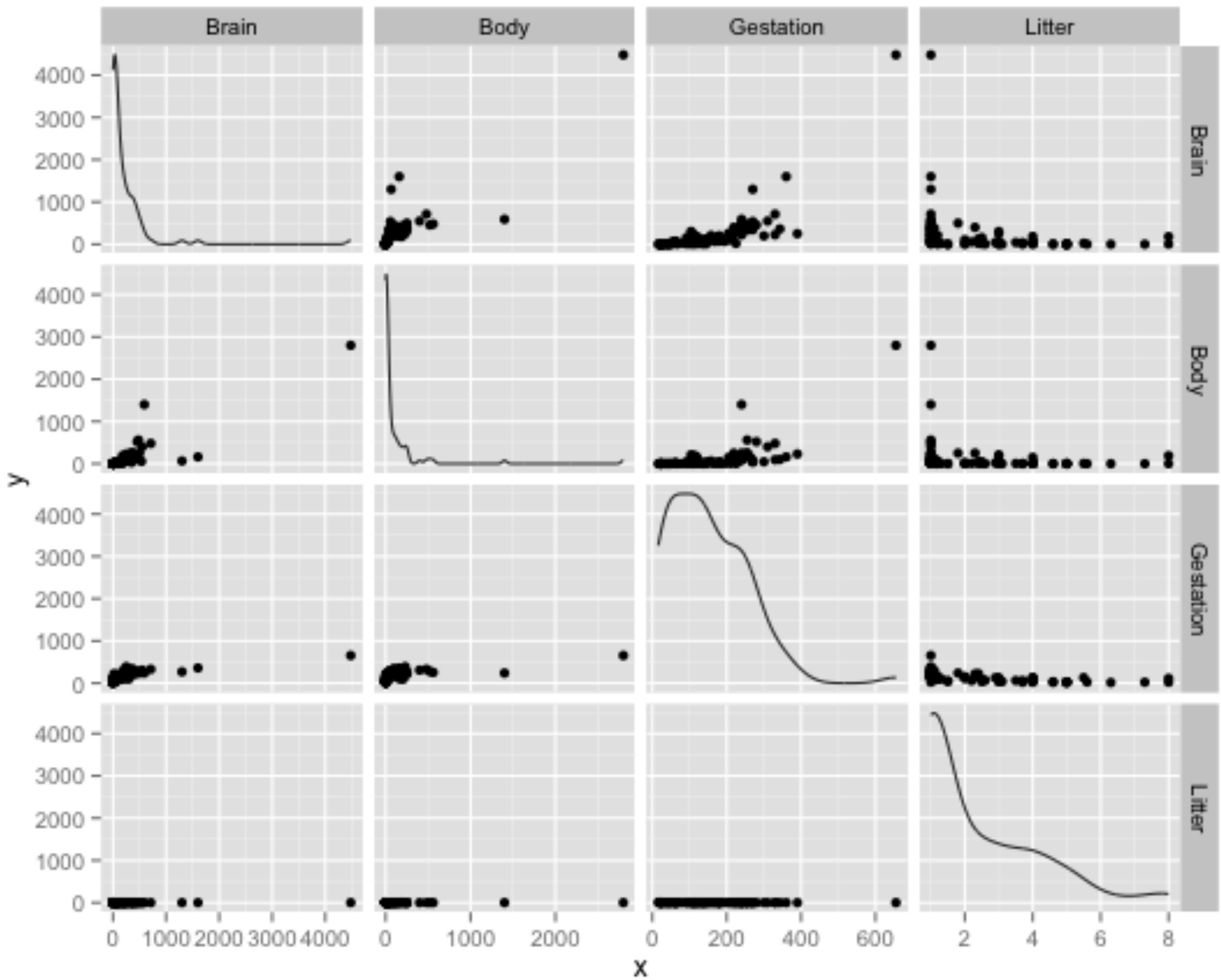
```
head(case0902)
            Species Brain  Body Gestation Litter
1            Quokka 17.50 3.500        26    1.0
2          Hedgehog  3.50 0.930        34    4.6
3        Tree shrew  3.15 0.150        46    3.0
4  Elephant shrew I  1.14 0.049        51    1.5
```

Scatterplot matrix

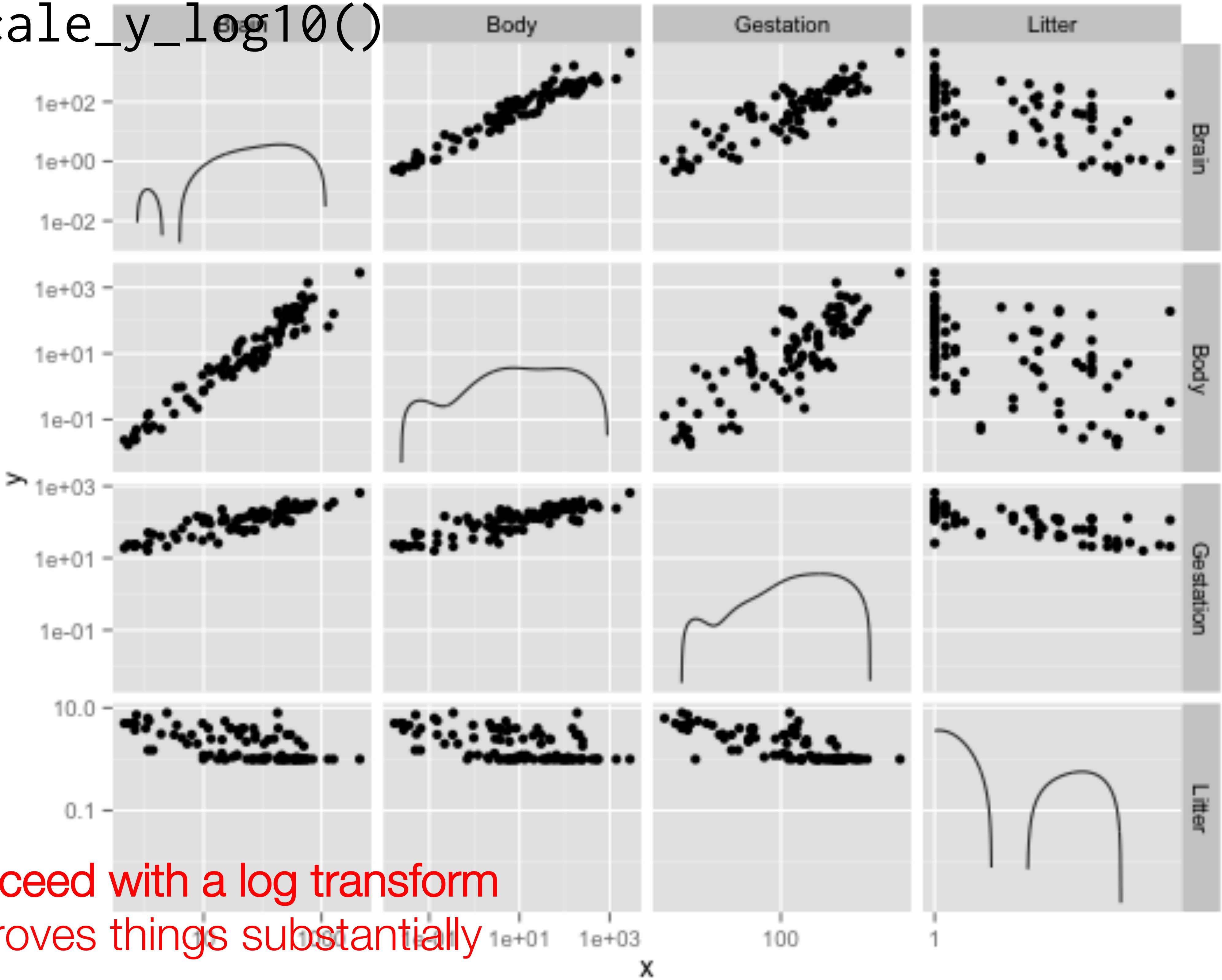
all pairwise scatterplots

```
plotmatrix(case0902[, -1])
```

not the first column

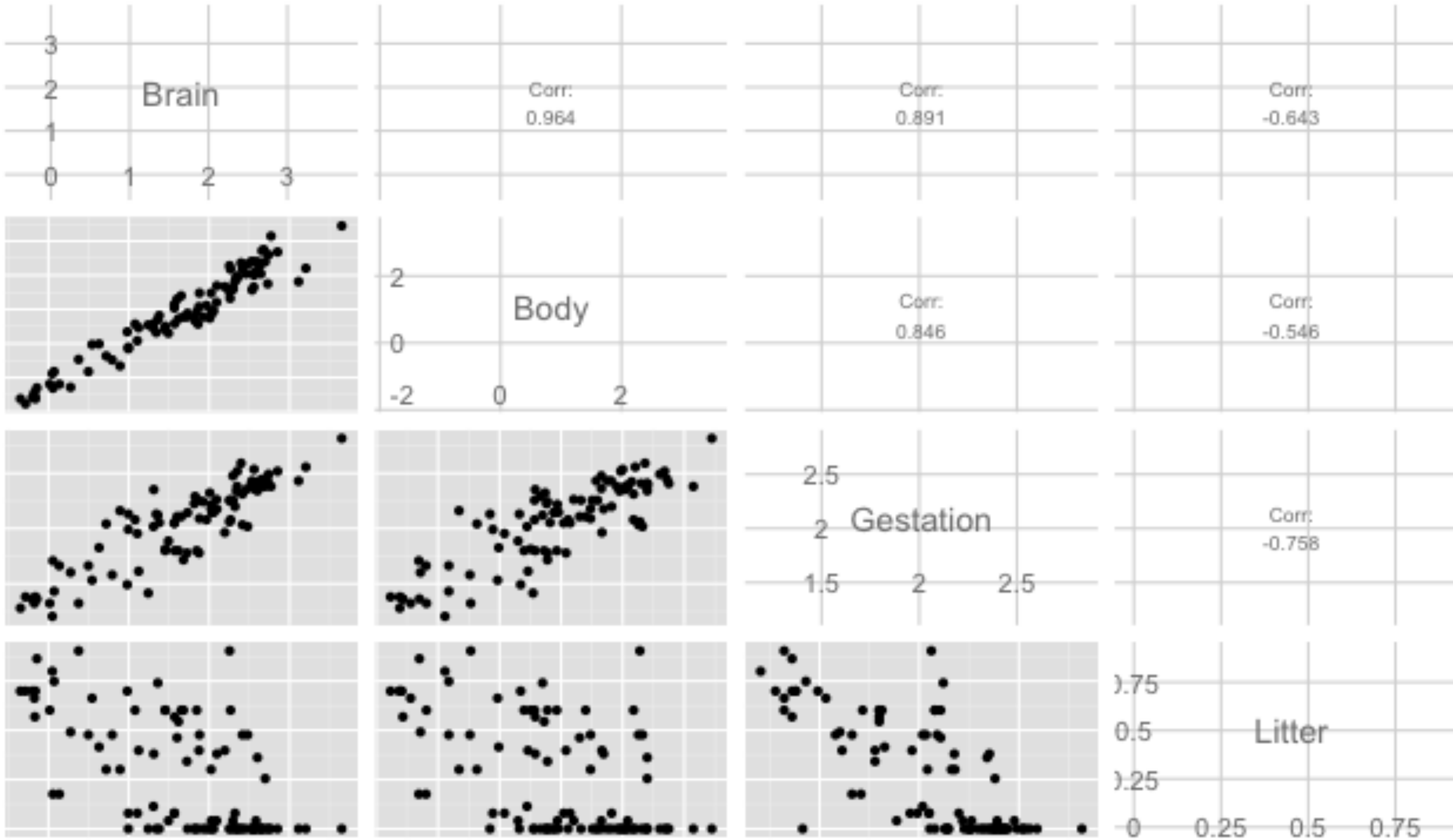


```
plotmatrix(case0902[, -1]) +
    scale_x_log10() +
    scale_y_log10()
```



Only proceed with a log transform
If it improves things substantially

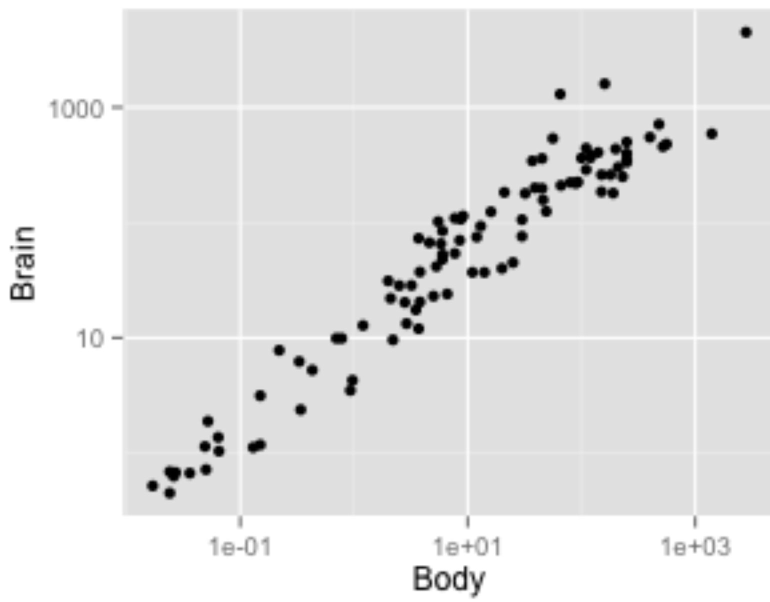
```
library(GGally)
# to log transform need to do each column
library(plyr)
case0902log <- colwise(log10, is.numeric)(case0902)
case0902log$Species <- case0902$Species
ggpairs(case0902log,columns = c(1:4))
```



better version

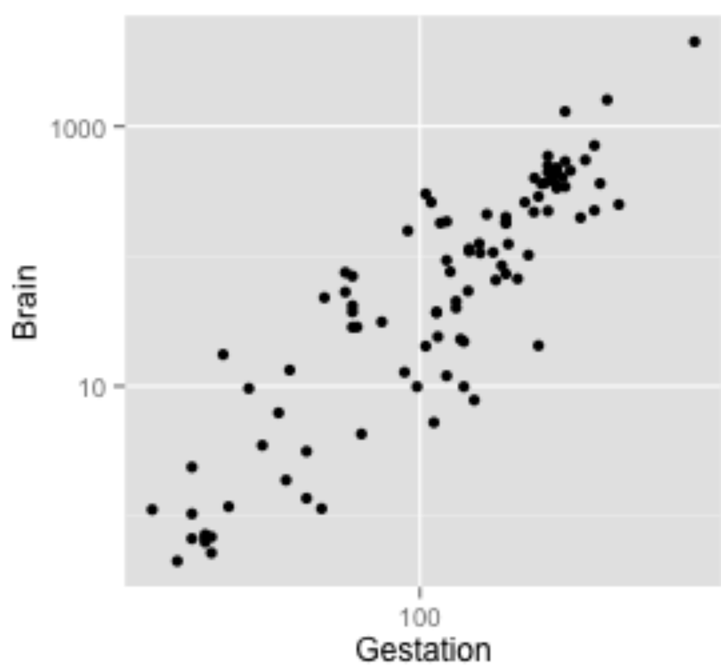
Or explore "by hand"

```
qplot(Body, Brain, data = case0902,
   log = "xy")
```



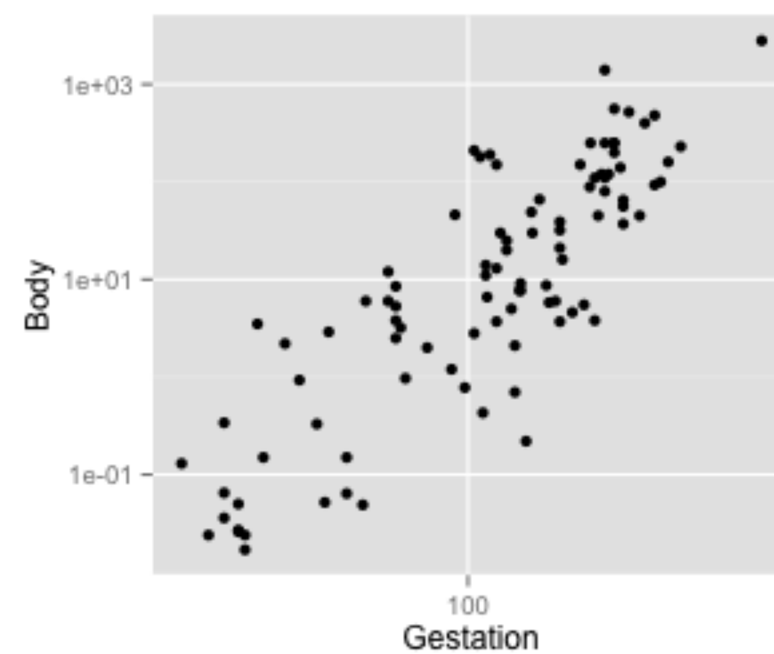
Positive correlation between brain weight and body weight

```
qplot( Gestation, Brain, data = case0902 ,
   log = "xy")
```



Positive correlation between Gestation length and brain weight

But maybe that is because there is a relationship between body weight and gestation length.




```
qplot(Gestation, Body,
    data = case0902,
    log = "xy")
```

Similarly for litter size

Your turn

$\mu\{\log(\text{brain}) \mid \text{gestation, body, litter}\} =$

$$\beta_0 + \beta_1 \log(\text{body}) + \beta_2 \log(\text{gestation})$$

What is the effect of log(gestation)?

How would we interpret β_2?

Interpretation depends on what else is in the model

The interpretation of β_1 is different in the two models:

1: $\mu\{\text{brain} \mid \text{gestation}\} = \beta_0 + \beta_1 \text{gestation}$

2: $\mu\{\text{brain} \mid \text{gestation, body}\} = \beta_0 + \beta_1 \text{gestation} + \beta_2 \text{body}$

1: β_1 is the rate of change of brain weight with changes in gestation length, over all mammals.

2: β_1 is the rate of change of brain weight with changes in gestation length, holding body size fixed (or within mammals of the same body size).

β_1 in 1 could be non-zero, because brain weight and gestation length are associated, or because both brain weight and gestation length are associated with body size.

A tentative model

$\mu\{\log(\text{brain}) \mid \text{gestation, body, litter}\} =$

$$\beta_0 + \beta_1 \log(\text{body}) + \beta_2 \log(\text{gestation}) + \beta_3 \log(\text{litter})$$

We know brain weight is related to body size, so we need the β_1 term in the model.

If both β_2 and $\beta_3 = 0$, then neither are associated with brain size after accounting for body size.

If $\beta_2 \neq 0$ then brain size is related to gestation length after accounting for body size and litter size.

If $\beta_3 \neq 0$ then brain size is related to litter size after accounting for body size and gestation.

Shorthand: $\mu\{\log(\text{brain}) \mid \text{gestation, body, litter}\} = \log(\text{body}) + \log(\text{gestation}) + \log(\text{litter})$

```
> summary(lm(log(Brain) ~ log(Body) + log(Gestation) + log(Litter),
         data = case0902))

...

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.85482    0.66167   1.292  0.19962
log(Body)       0.57507    0.03259  17.647  < 2e-16 ***
log(Gestation)  0.41794    0.14078   2.969  0.00381 **
log(Litter)    -0.31007    0.11593  -2.675  0.00885 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There was strong evidence that brain weight was associated with either gestation length or litter size, even after accounting for the effect of body weight. (not in this output!)

There was strong evidence that litter size was associated with brain weight after accounting for body weight and gestation (p-value = 0.0089).

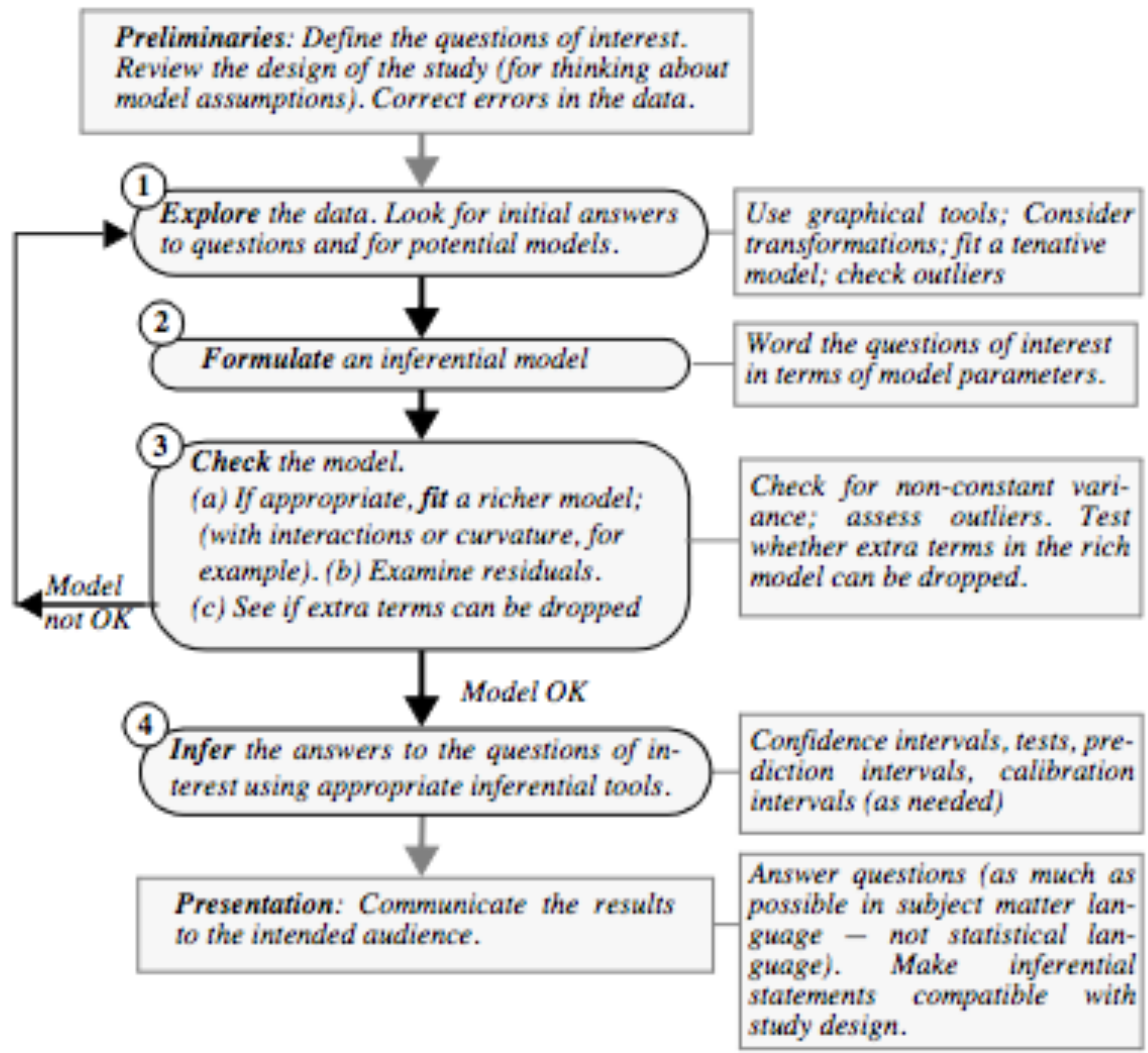
There was strong evidence that gestation length was associated with brain weight after accounting for body weight and litter size (p-value = 0.0038).

Observational study

Strategy

Display 9.9 **p. 251**

A strategy for data analysis using statistical models



The first thing you need to consider, is:

Will my regression model answer my questions of interest?

Steps 1 & 2

The second:

Is my regression model an appropriate model for my data?

Steps 1 & 3

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.

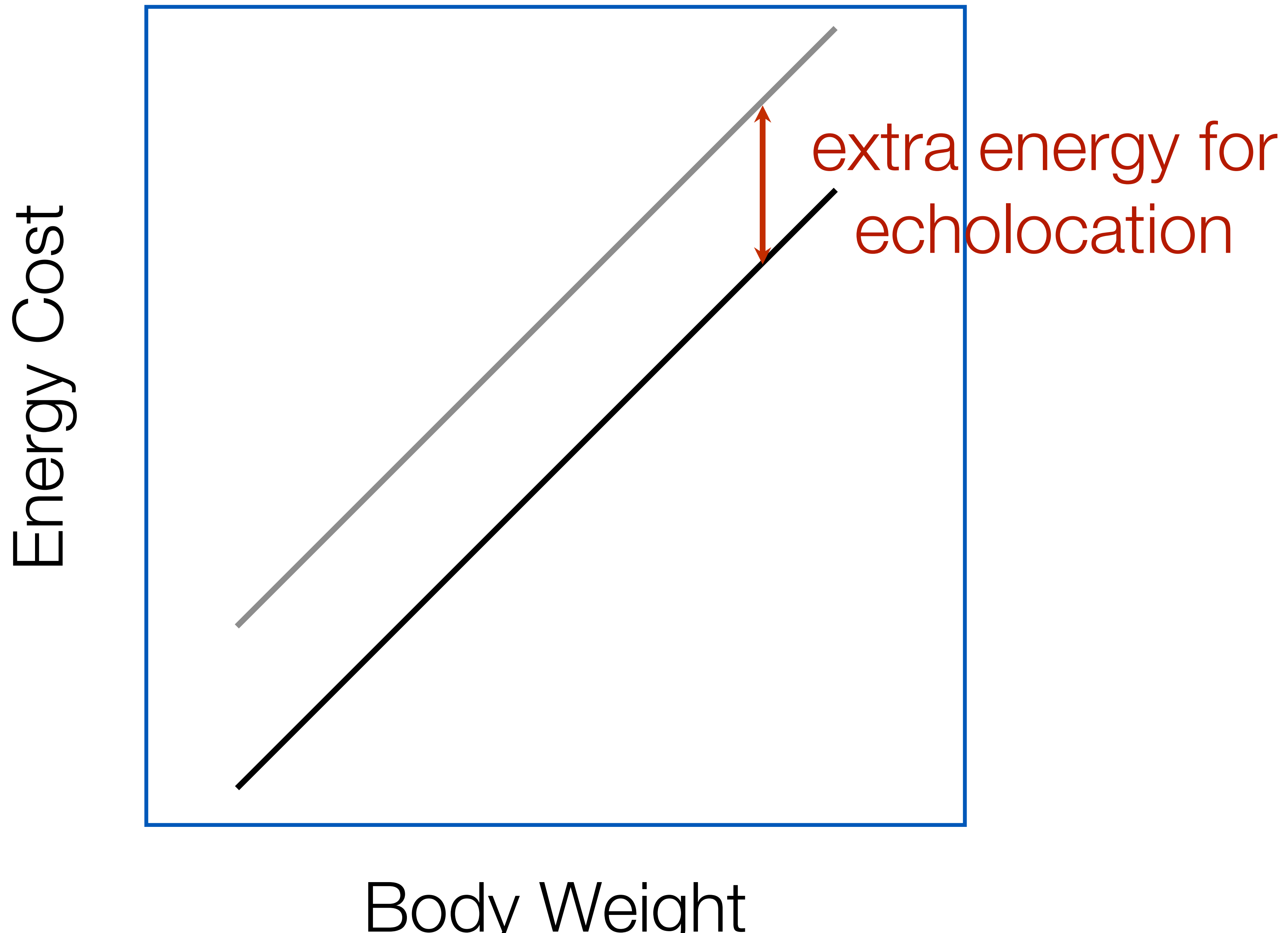
Zoologists wonder whether the energy costs of echolocation during flight are the sum of flights costs plus echolocation.

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

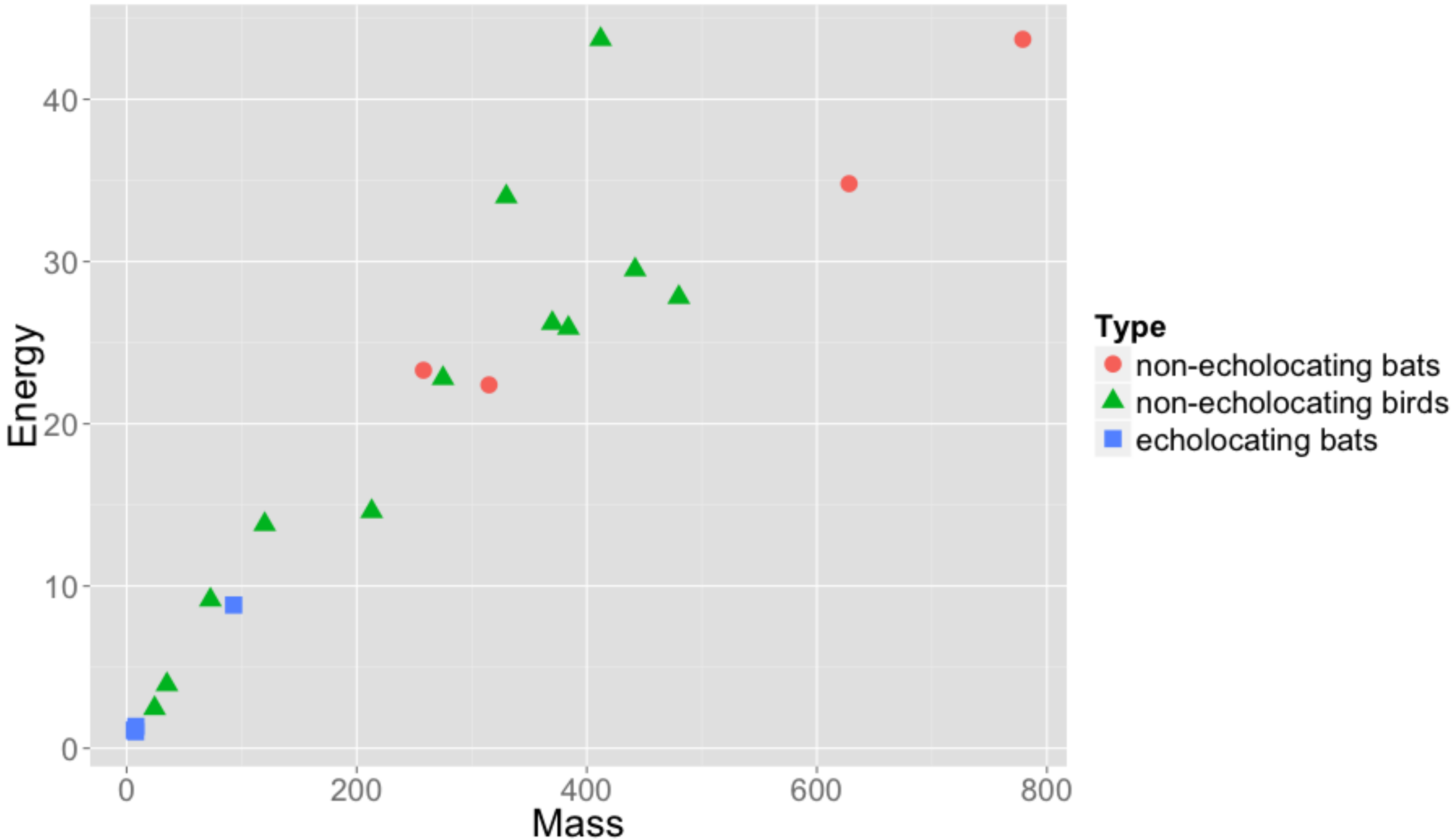
Complication: the energy costs of flight depend on how heavy you are

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.



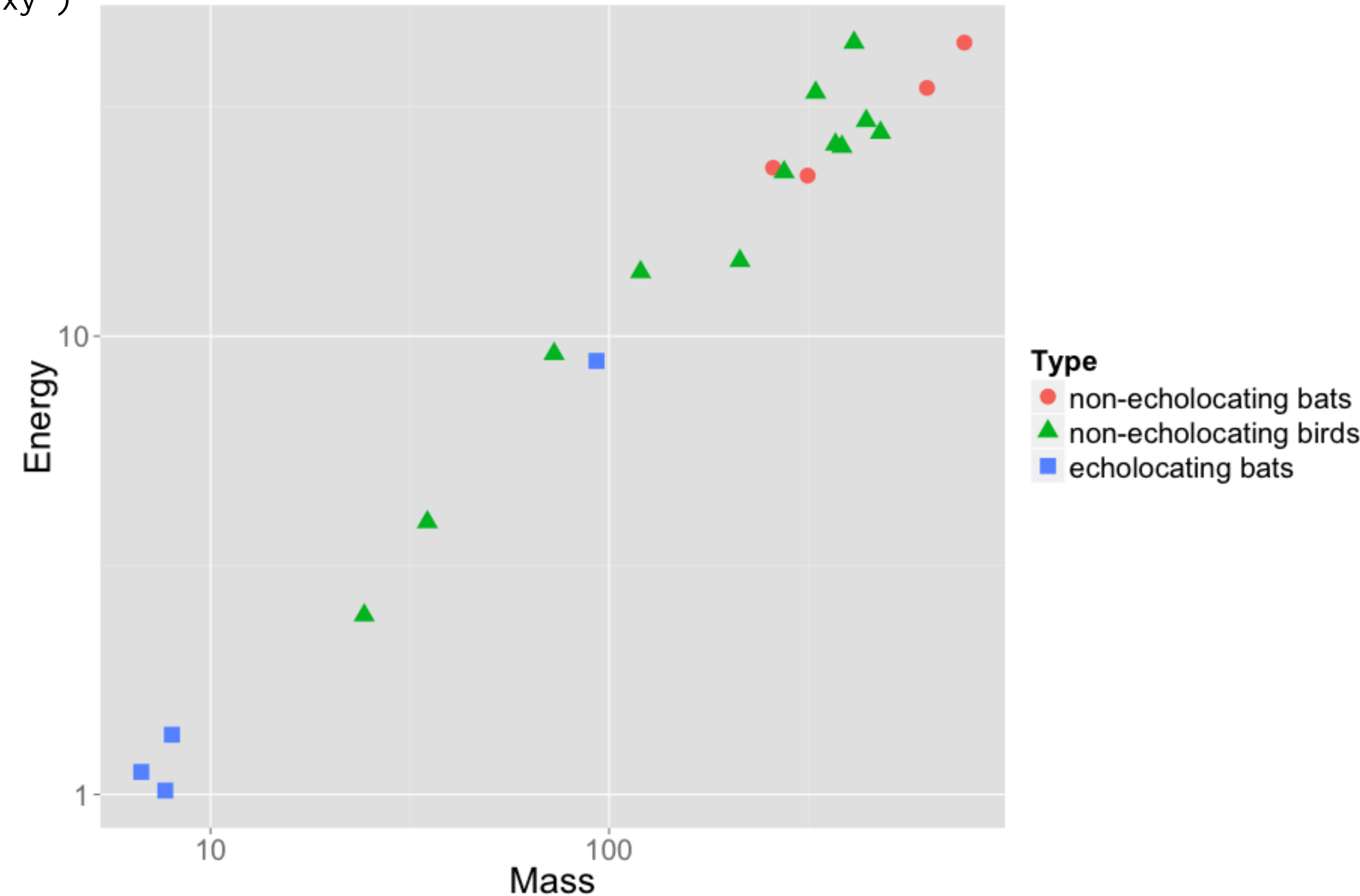
```
qplot(Mass, Energy, data = case1002, colour = Type, shape = Type)
```



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

μ{ 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

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

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

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:

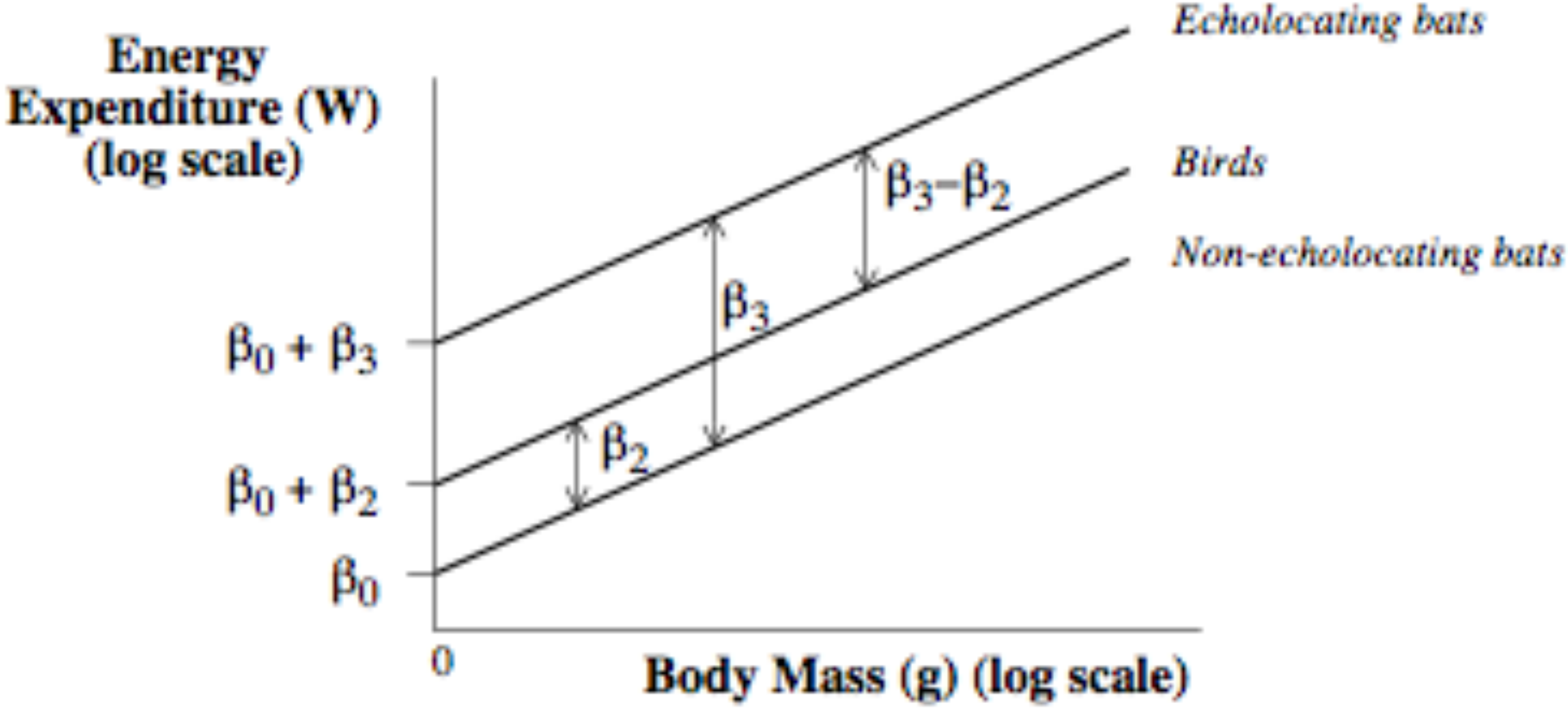
$\mu\{\log \text{Energy} \mid \log \text{Mass}, \text{ebat} = 0, \text{bird} = 1\} =$

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

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, $\beta_3 = 0$.

Inference on a single parameter, today

Is the model appropriate for our data?

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

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

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

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

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

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).

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

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

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

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

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 $\hat{\sigma}$.

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