Stat 412/512

Another multiple Regression

Jan 14 2015

Charlotte Wickham

stat512.cwick.co.nz

Your turn

Consider the model:

$\mu\{ \textit{flowers} \mid \textit{light}, \textit{early}\} =$

$\beta_0 + \beta_1 \textit{light} + \beta_2 \textit{early} + \beta_3(\textit{light} \times \textit{early})$

What is the mean flowers per plant for units in the late treatment group?

$\beta_0 + \beta_1 \text{light}$

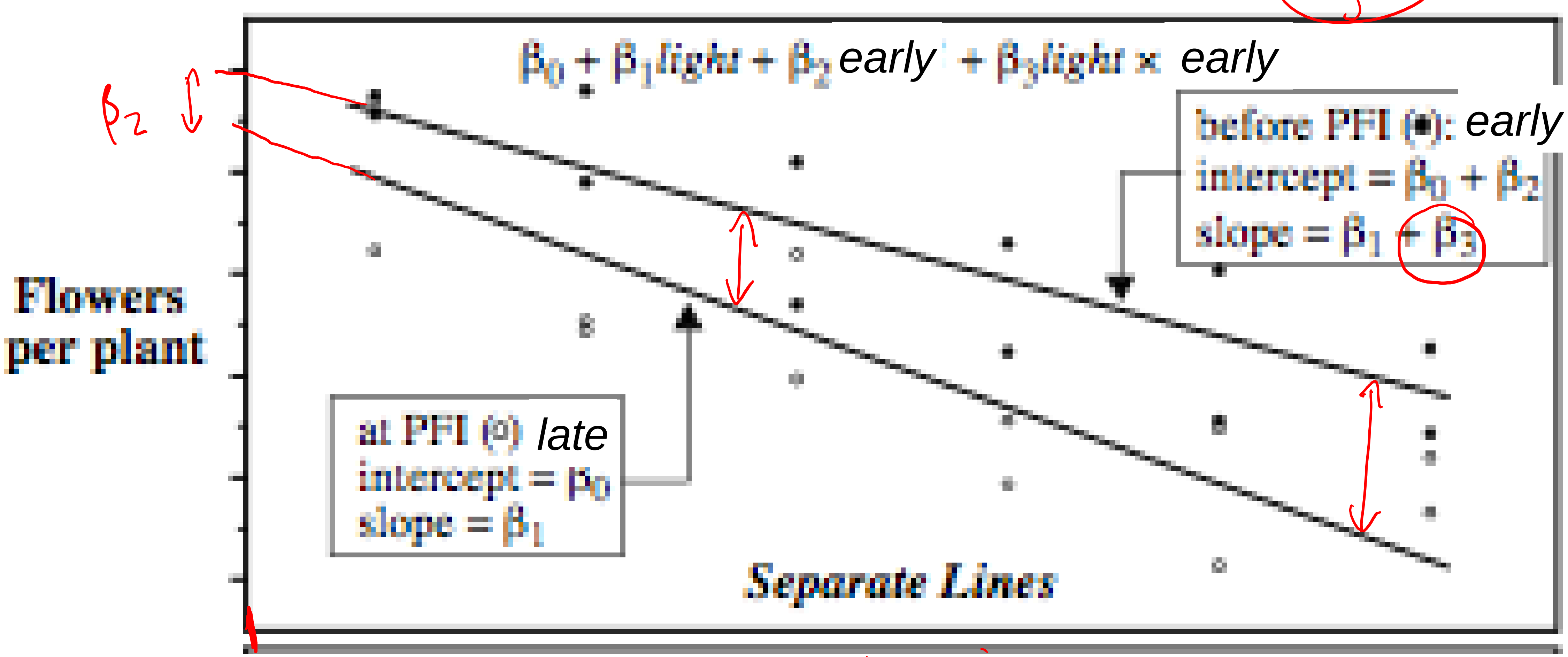
What is the mean flowers per plant for units in the early treatment group?

$(\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{light}$

Separate lines model

$\mu\{\textit{flowers} \mid \textit{light}, \textit{early}\} =$

$\beta_0 + \beta_1 \textit{light} + \beta_2 \textit{early} + \beta_3(\textit{light} \times \textit{early})$



The effect of light intensity depends on timing

Interaction terms

Two variables are said to **interact** if the effect of one variable on the mean response depends on the other variable.

β_3(*light* × *early)* is called an **interaction** term. In our example it allows the effect of intensity on mean number of flowers to depend on whether the timing was early or late. In this example, it allowed the mean for the *early* units to have a different slope with respect to *light* from the *late* units.

I.e. it allows *light* and *early* to interact.

Does the effect of the intensity depend on the timing of light treatment?

Parallel lines: the effect of light intensity doesn't depend on timing,

$\mu\{$ *flowers* | *light*, *early*$\}$ = $\beta_0 + \beta_1 light + \beta_2 early$

Separate lines: the effect of light intensity depends on timing

$\mu\{$ *flowers* | *light*, *early*$\}$ =

$\beta_0 + \beta_1 light + \beta_2 early + \beta_3(light \times early)$

What's the difference?

If $\beta_3 = 0$, the separate lines model reduces to the parallel lines model.

So, to answer our question, we could use the separate lines model and ask is $\beta_3 = 0$?

"...questions of interest are translated to statements about parameters."

separate lines model

```
> fit_sep <- lm(Flowers ~ Intens + early + I(Intens * early), data =
case0901)
> summary(fit_sep)

Call:
lm(formula = Flowers ~ Intens + early + I(Intens * early), data = case09

Residuals:
   Min     1Q Median     3Q    Max
-9.516 -4.276 -1.422  5.473 11.938

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
β0 (Intercept)   71.623333   4.343305  16.491 4.14e-13 ***
β1 Intens        -0.041076   0.007435  -5.525 2.08e-05 ***
β2 early         11.523333   6.142361   1.876   0.0753 .
β3 I(Intens * early)  0.001210   0.010515   0.115   0.9096
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.598 on 20 degrees of freedom
Multiple R-squared: 0.7993, Adjusted R-squared: 0.7692
F-statistic: 26.55 on 3 and 20 DF,  p-value: 3.549e-07
```

$\beta_j = 0$

There is no evidence that the effect of Intensity depends on timing.

parallel lines model

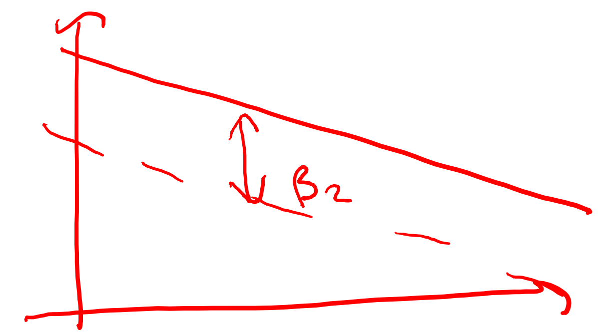
```
> fit_par <- lm(Flowers ~ Intens + early, data = case0901)
> summary(fit_par)

Call:
lm(formula = Flowers ~ Intens + early, data = case0901)

Residuals:
   Min      1Q  Median      3Q     Max
-9.652 -4.139 -1.558  5.632 12.165

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
β0 (Intercept) 71.305834   3.273772  21.781 6.77e-16 ***
β1 Intens      -0.040471   0.005132  -7.886 1.04e-07 ***
β2 early       12.158333   2.629557   4.624 0.000146 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.441 on 21 degrees of freedom
Multiple R-squared: 0.7992, Adjusted R-squared:  0.78
F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```



Increasing light intensity decreased the mean number of flowers per plant by 4.0 flowers for every 100 *μmol/m²/sec.*

Beginning the light treatments 24 days before PFI increased the mean number of flowers per plant by 12.1 compared to beginning light treatments at PFI.

$\beta_0 = 71.3$

The mean number of flowers per plant, is estimated to be 71.3 when the light intensity is 0 μmol/m²/sec and applied at PFI.

Today

A couple of points on constructed variables

Another example of multiple regression

Some new plotting methods

Indicators for more than two categories

A collection of indicator variables can be used for variables with more than two categories.

L300 could be an indicator for Intensity = 300.

L450 could be an indicator for Intensity = 450.

... and so on

$\mu\{\, flowers \mid light, early \} = \beta_0 + \beta_1 L300 + \beta_2 L450 +$

$+ \beta_3 L600 + \beta_4 L750 + \beta_5 L900 + \beta_2 early$

Your turn

$\mu\{ \textit{flowers} \mid \textit{light}, \textit{early}\} = \beta_0 + \beta_1 L300 + \beta_2 L450 +$

$+ \beta_3 L600 + \beta_4 L750 + \beta_5 L900 + \beta_6 \textit{early}$

What's the mean number of flowers when intensity is 300?

intensity = 300, L300 = 1

$\beta_0 + \beta_1 + \beta_6 \textit{early}$

What's the mean number of flowers when intensity is 150?

$\beta_0 + \beta_6 \textit{early}$

To fully represent I categories you need I-1 indicator variables.

The category without an indicator variable, becomes the baseline category.

A parameter (β) for an indicator variable, gives that level it's own intercept, and the parameter describes the difference between the intercept for that level and the baseline level.

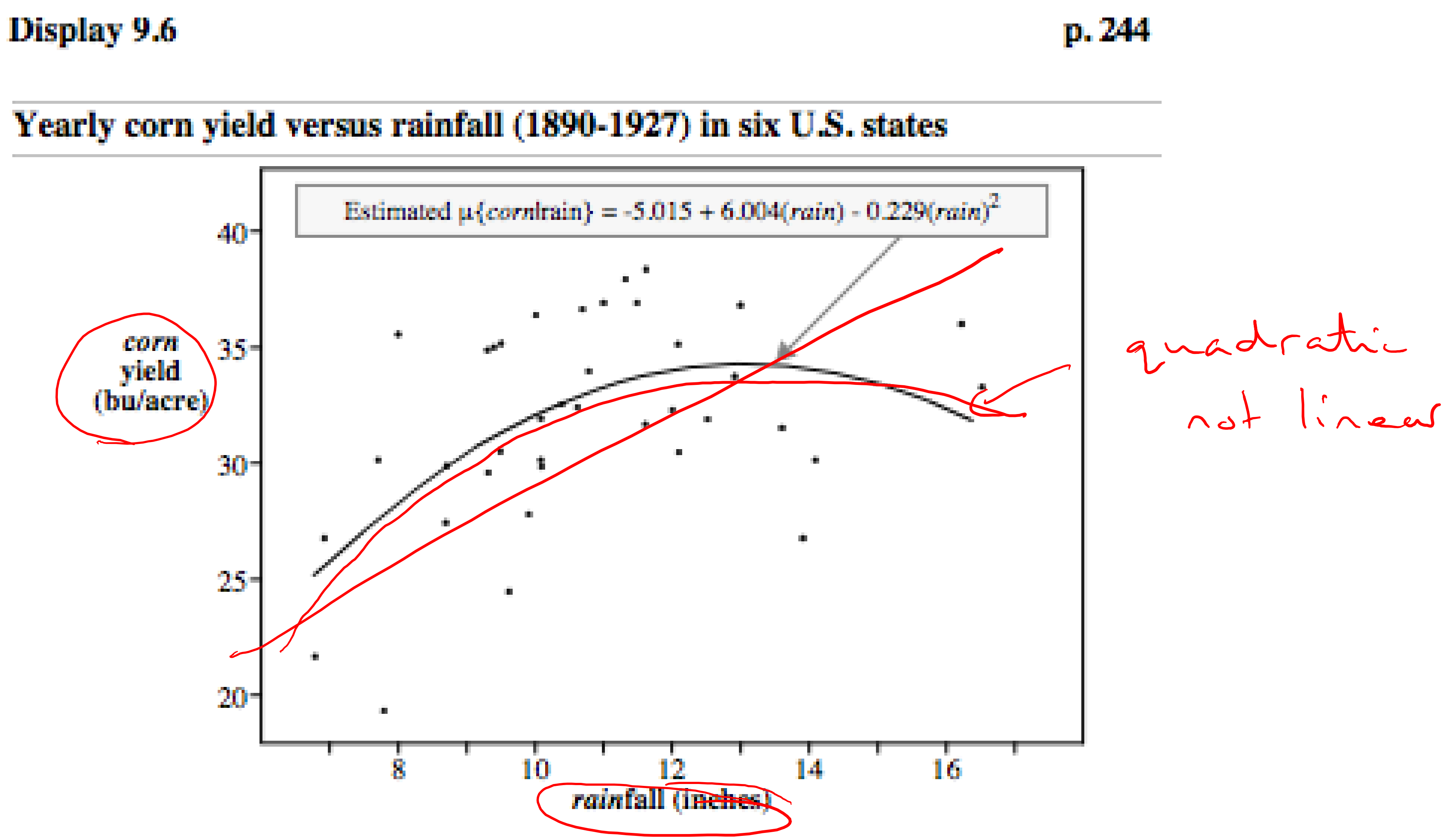
β_2 L300

A parameter (β) for an interaction between an indicator and another variable, gives that level it's own slope (w.r.t to the interacting variable) and the parameter describes the difference between the slope for that level and the slope for the baseline level.

β_7 L300 x early

If in doubt: work out the models for the mean for each category.

Squared terms for curvature



$$\mu\{\textit{corn yield} \mid \textit{rainfall}\} = \beta_0 + \beta_1 \textit{rainfall} + \beta_2 \textit{rainfall}^2$$

Shorthand

Shorthand: UPPERCASE for indicator variables,
leave out parameters

$\mu\{\,\textit{flowers} \mid \textit{Intensity}, \textit{early}\,\} = \text{INTENSITY} + \textit{early}$

$= \beta_0 + \beta_1 L300 + \beta_2 L450 + \ldots + \beta_6\,\text{early}$

$\mu\{\,\textit{flowers} \mid \textit{Intensity}, \textit{Time}\,\} = \text{INTENSITY} + \textit{TIME}$

Later
early

$\mu\{\,\textit{flowers} \mid \textit{Intensity}, \textit{Time}\,\} = \text{Intensity} + \textit{TIME}$

$\beta_0 + \beta_1\,\text{intensity} + \beta_2\,\text{early}$

$\mu\{\,\textit{flowers} \mid \textit{Intensity}, \textit{Time}\,\} = \text{Intensity} + \textit{TIME} + (\textit{Intensity} \times \textit{TIME})$

$\mu\{\,\textit{corn yield} \mid \textit{rainfall}\,\} = \textit{rainfall} + \textit{rainfall}^2$

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
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 Elep	[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

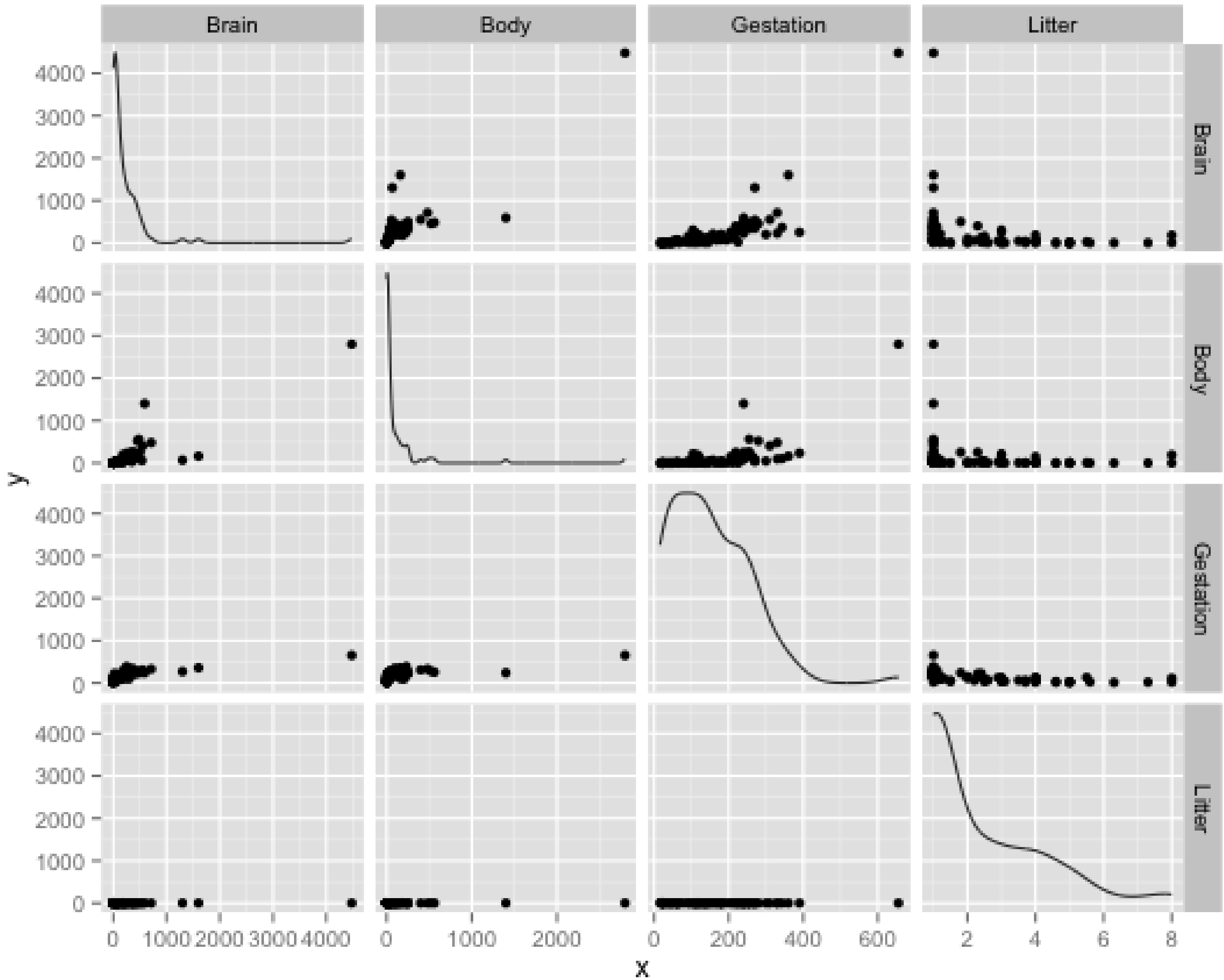
```
head(case0902)
            Species Brain  Body Gestation Litter
1            Quokka 17.50 3.500        26    1.0
2          Hedgehog  3.50 0.930        34    4.6
3        Tree shrew  3.15 0.150        46    3.0
4  Elephant shrew I  1.14 0.049        51    1.5
```

Scatterplot matrix

all pairwise scatterplots

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

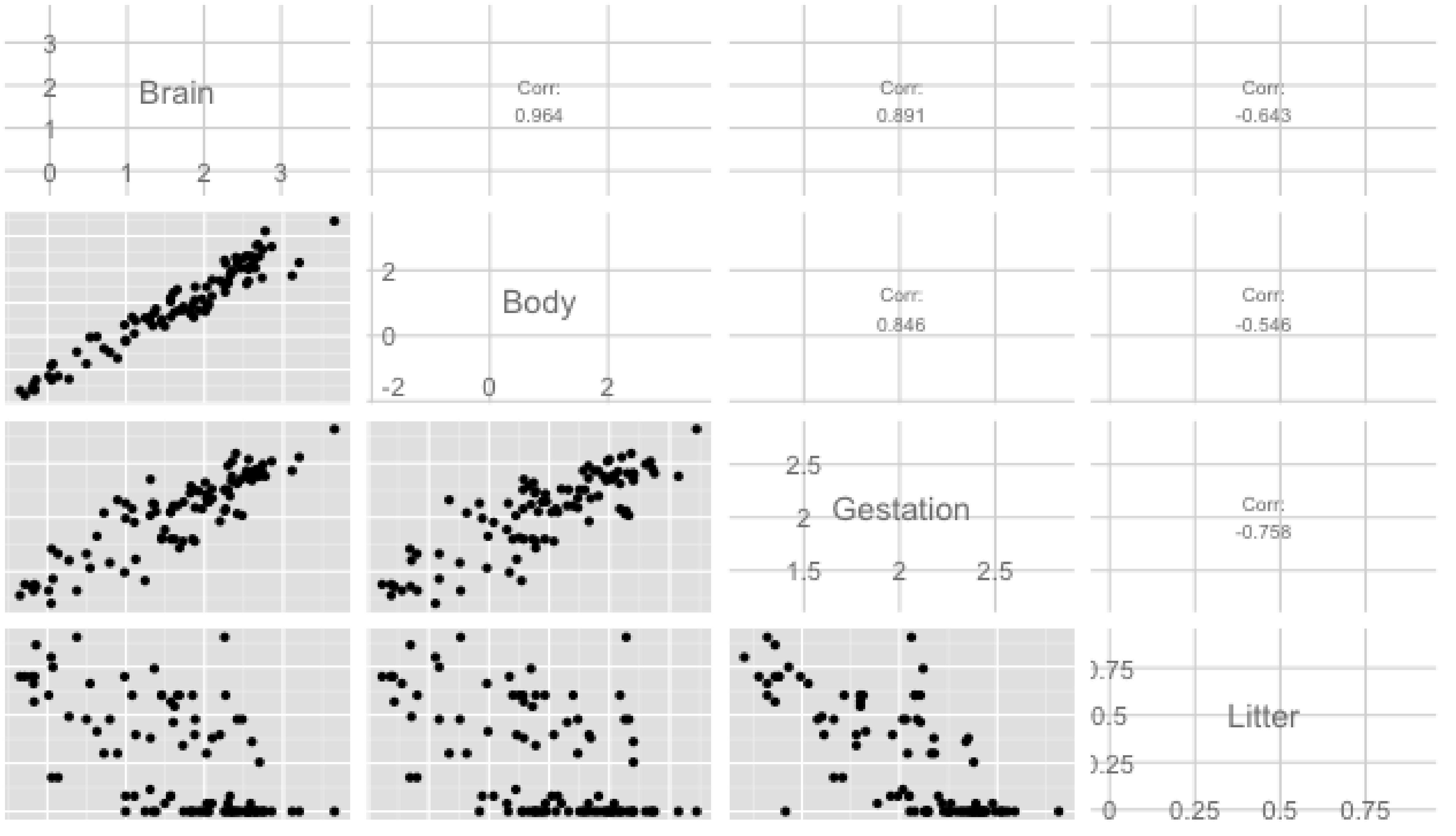
not the first column



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

Don't use log unless it helps!

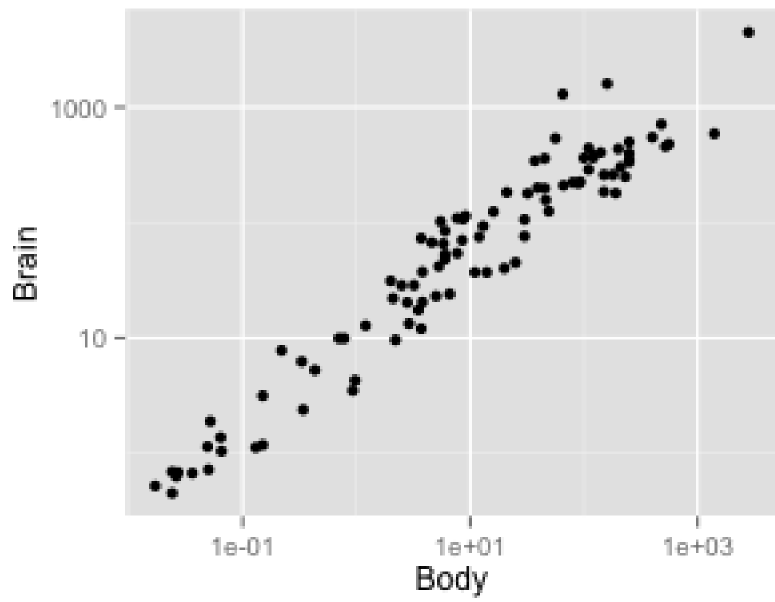
```
library(GGally)
# to log transform need to do each column
library(plyr)
case0902log <- colwise(log10, is.numeric)(case0902)
case0902log$Species <- case0902$Species
ggpairs(case0902log,columns = c(1:4))
```



better version

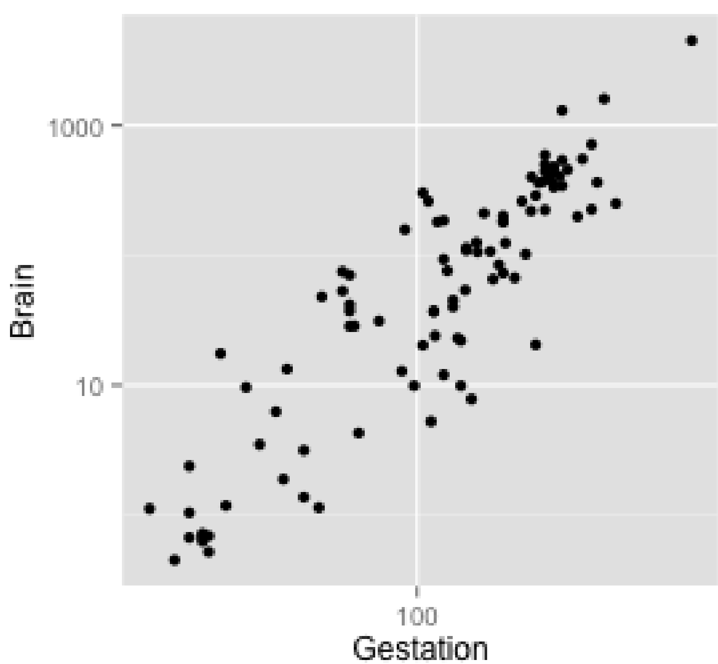
Or explore "by hand"

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



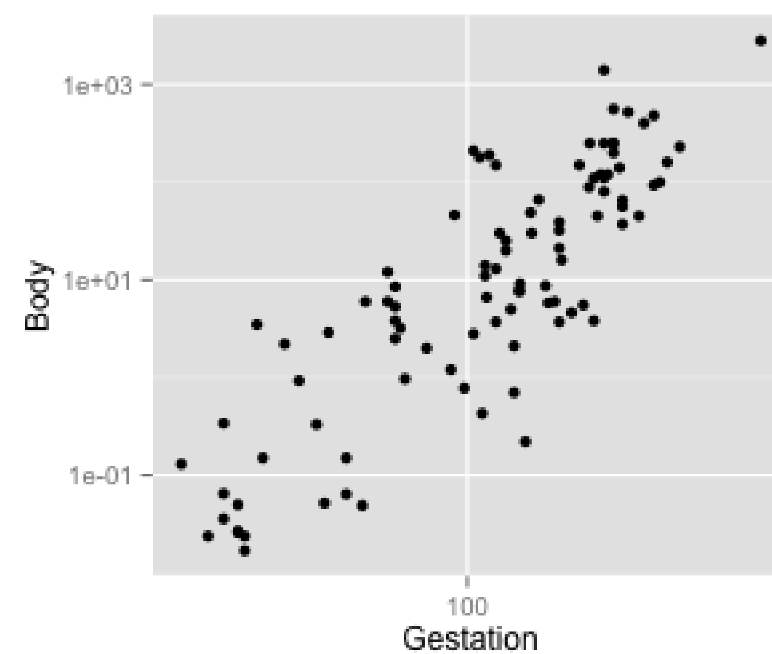
Positive correlation between brain weight and body weight

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 litte 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 fo 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