Multiple Regression EDA

David Gerard

2018-12-07

Learning Objectives

- A strategy for exploratory data analysis for multiple linear regression.
- Chapter 9.

Brain Size

- What variables are associated with brain weight?
- Collected information on 96 different species.
- We know that body weight is already associated with brain weight,
    - So what variables are associated with brain weight **after controlling for body weight**.
- Possible variables: Body weight (kg), gestation period (days), litter size

Brain Size

```
library(Sleuth3)
data("case0902")
head(case0902)
```

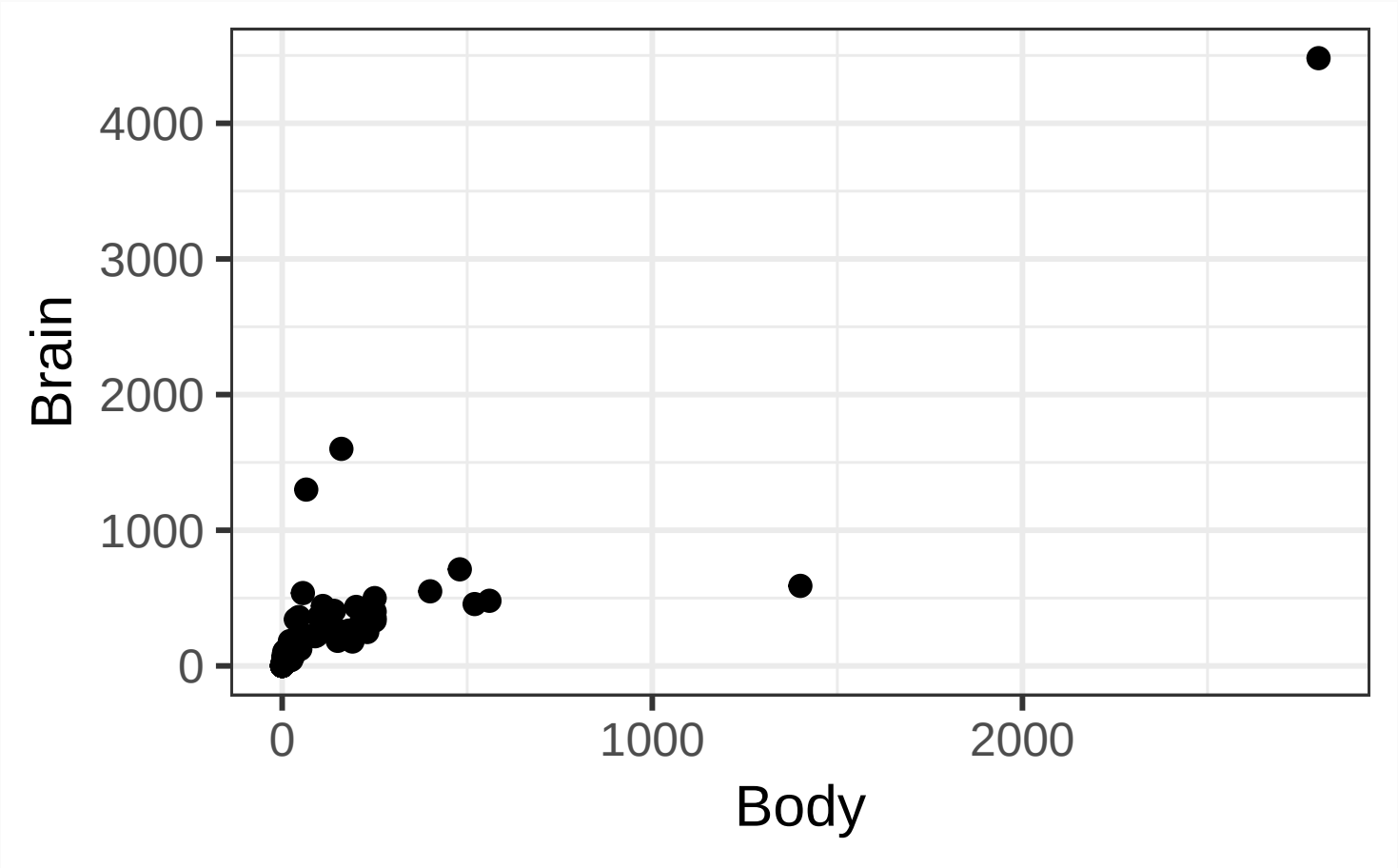
```
##              Species  Brain    Body Gestation Litter
## 1           Aardvark    9.6    2.20        31    5.0
## 2           Acouchis    9.9    0.78        98    1.2
## 3 African elephant 4480.0 2800.00       655    1.0
## 4           Agoutis   20.3    2.80       104    1.3
## 5         Axis deer  219.0   89.00       218    1.0
## 6            Badger   53.0    6.00        60    2.2
```

Scatterplots

- The first step is almost always **making a ton of scatterplots**.
- Plots of the response against each explanatory variable shows us what variables seems to be **marginally** related to the response.
- "Marginally related" = related **un**conditional on other variables.

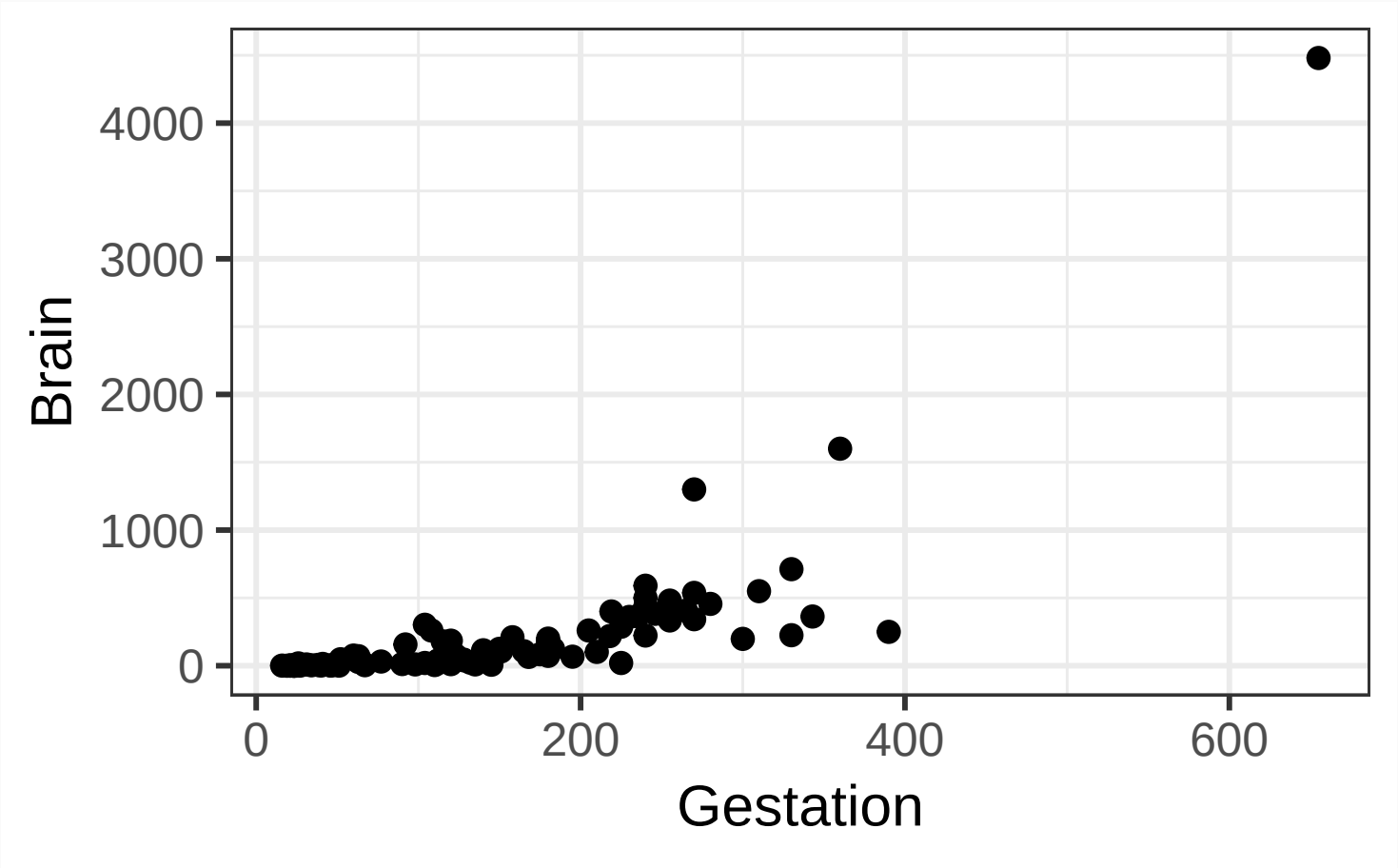
Response against each explanatory variable

```
qplot(Body, Brain, data = case0902)
```



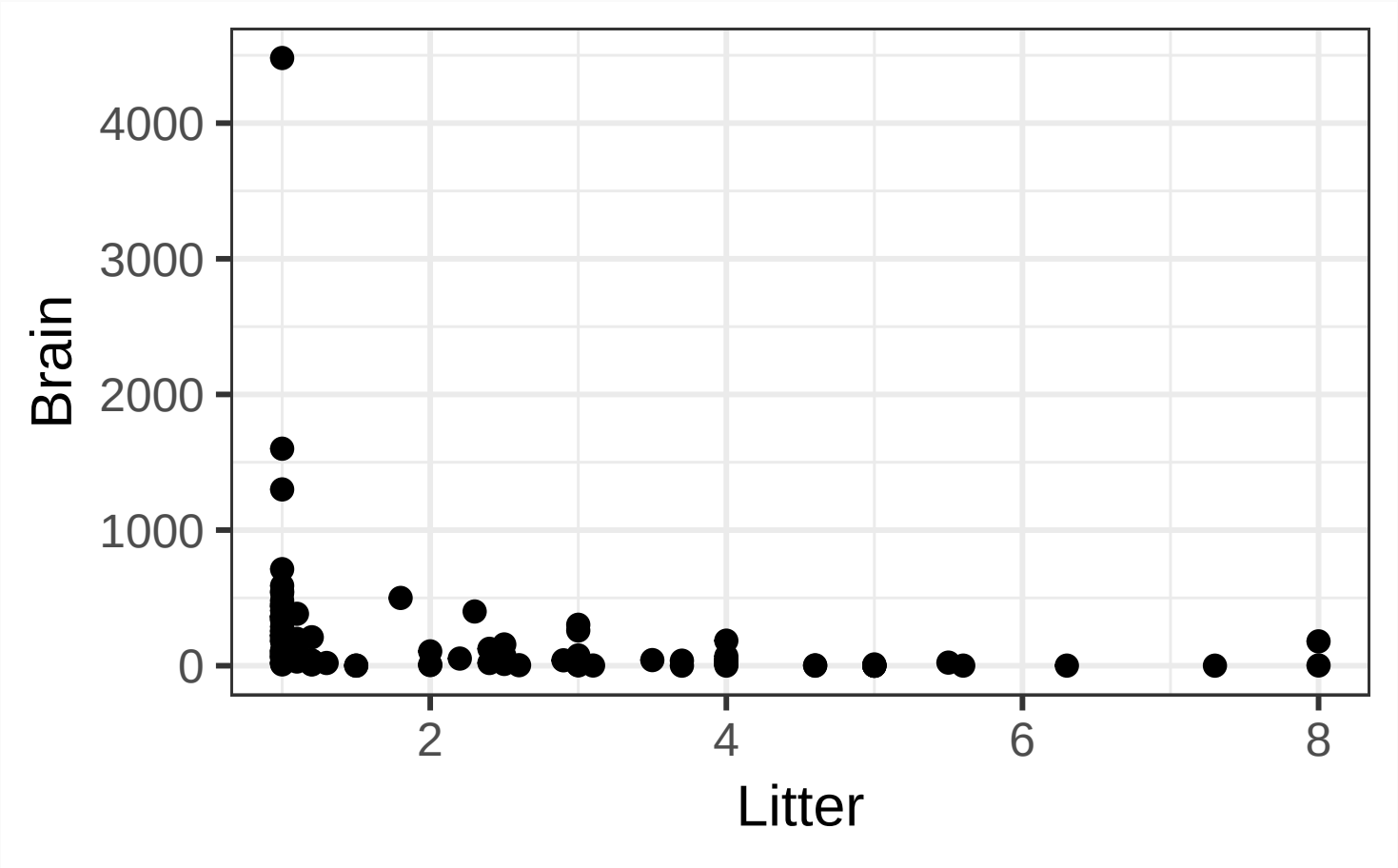
Response against each explanatory variable

```
qplot(Gestation, Brain, data = case0902)
```



Response against each explanatory variable

```
qplot(Litter, Brain, data = case0902)
```



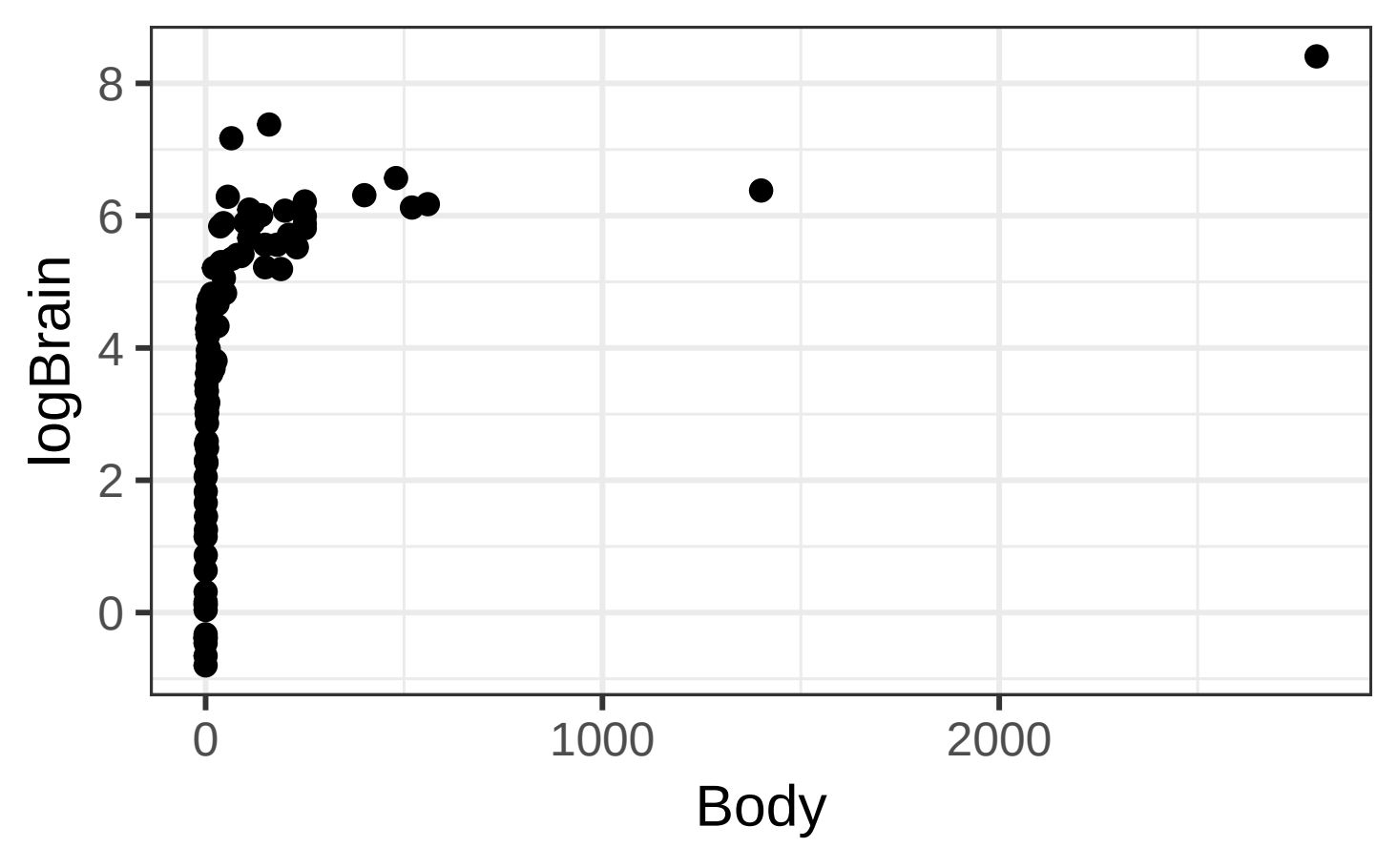
Response against each explanatory variable

- Curvature and different spreads at each explanatory variable suggest a log transformation of `Brain`.

```
case0902$logBrain <- log(case0902$Brain)
```

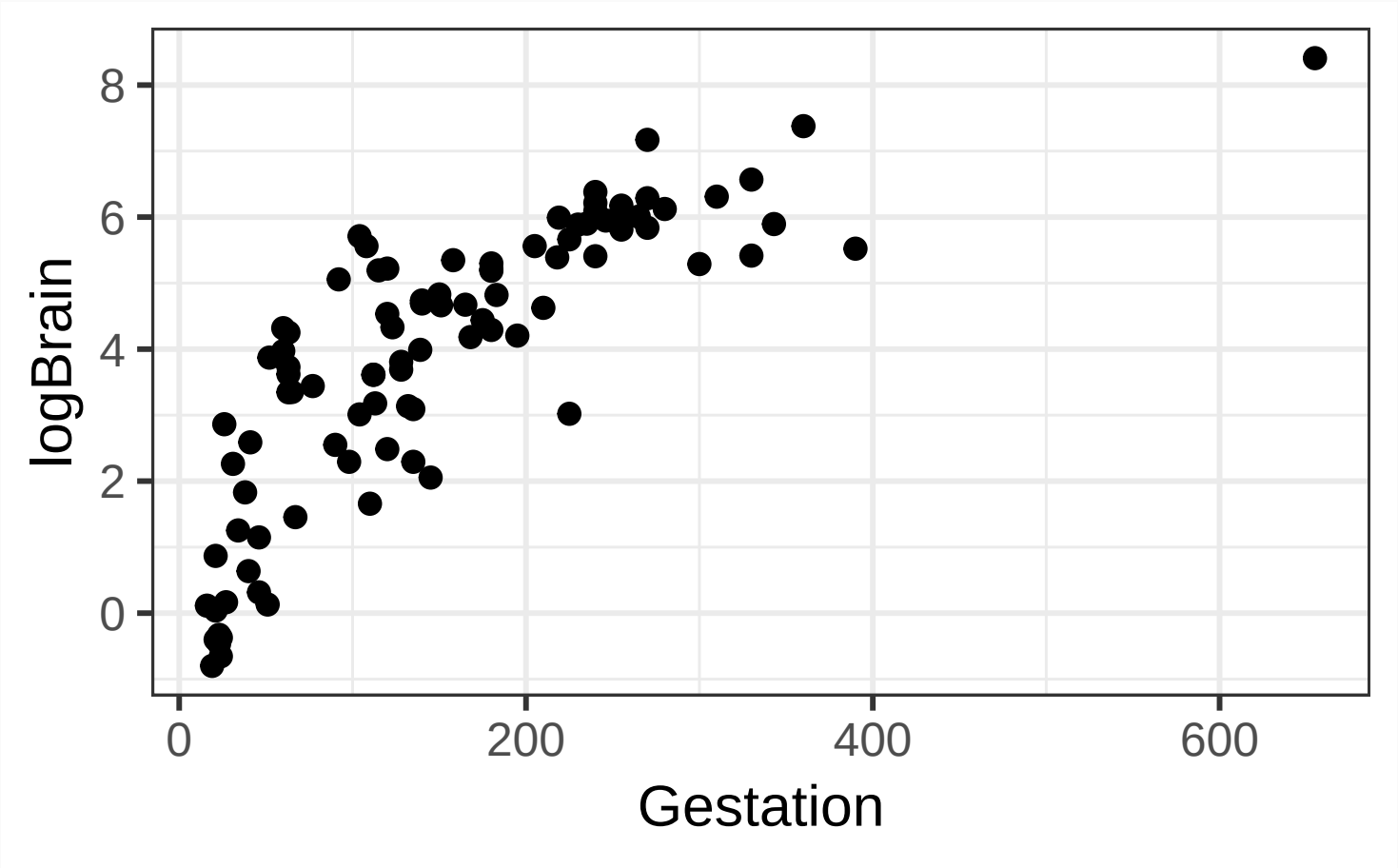
Response against each explanatoery variable

```
qplot(Body, logBrain, data = case0902)
```



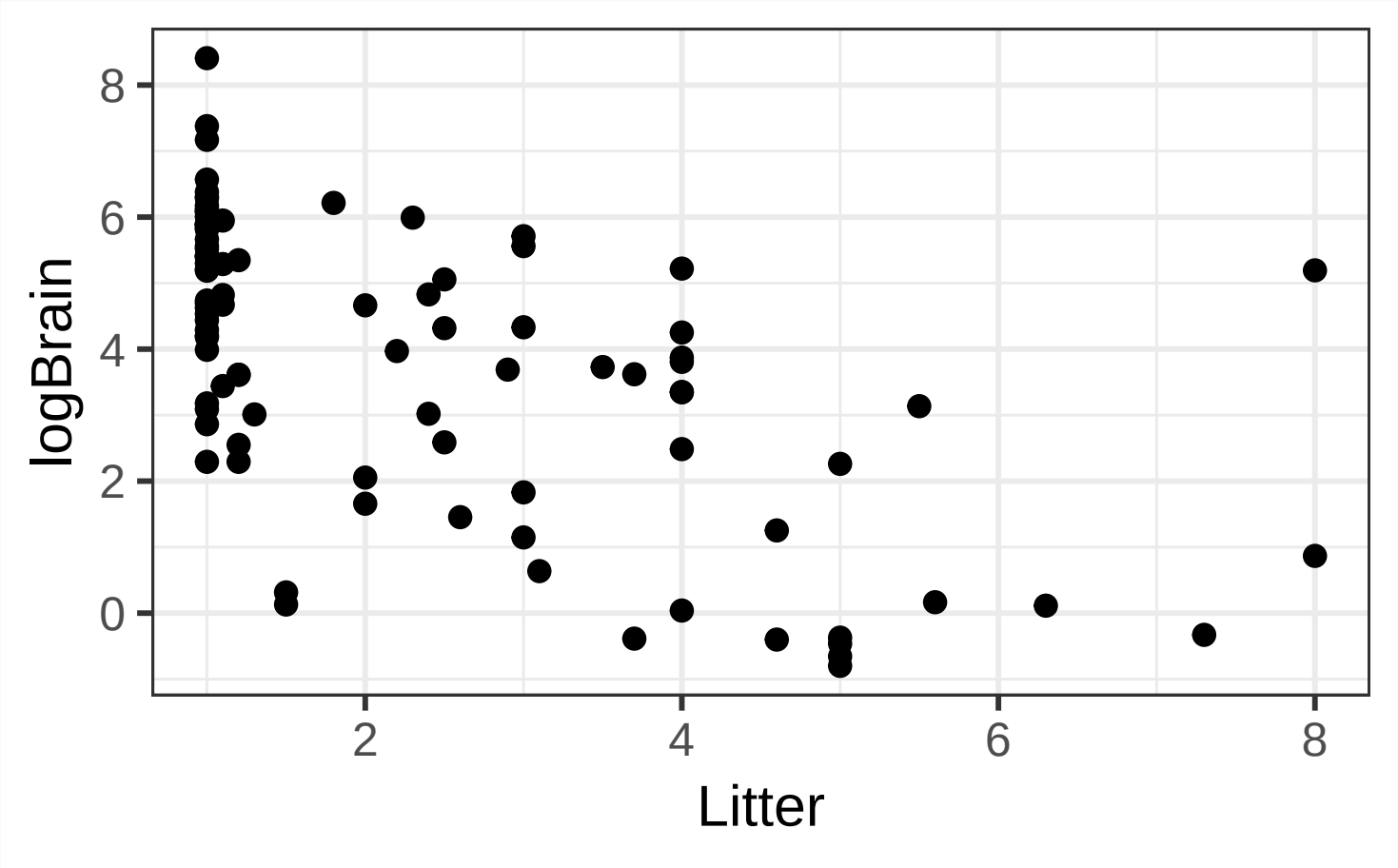
Response against each explanatory variable

```
qplot(Gestation, logBrain, data = case0902)
```



Response against each explanatory variable

```
qplot(Litter, logBrain, data = case0902)
```

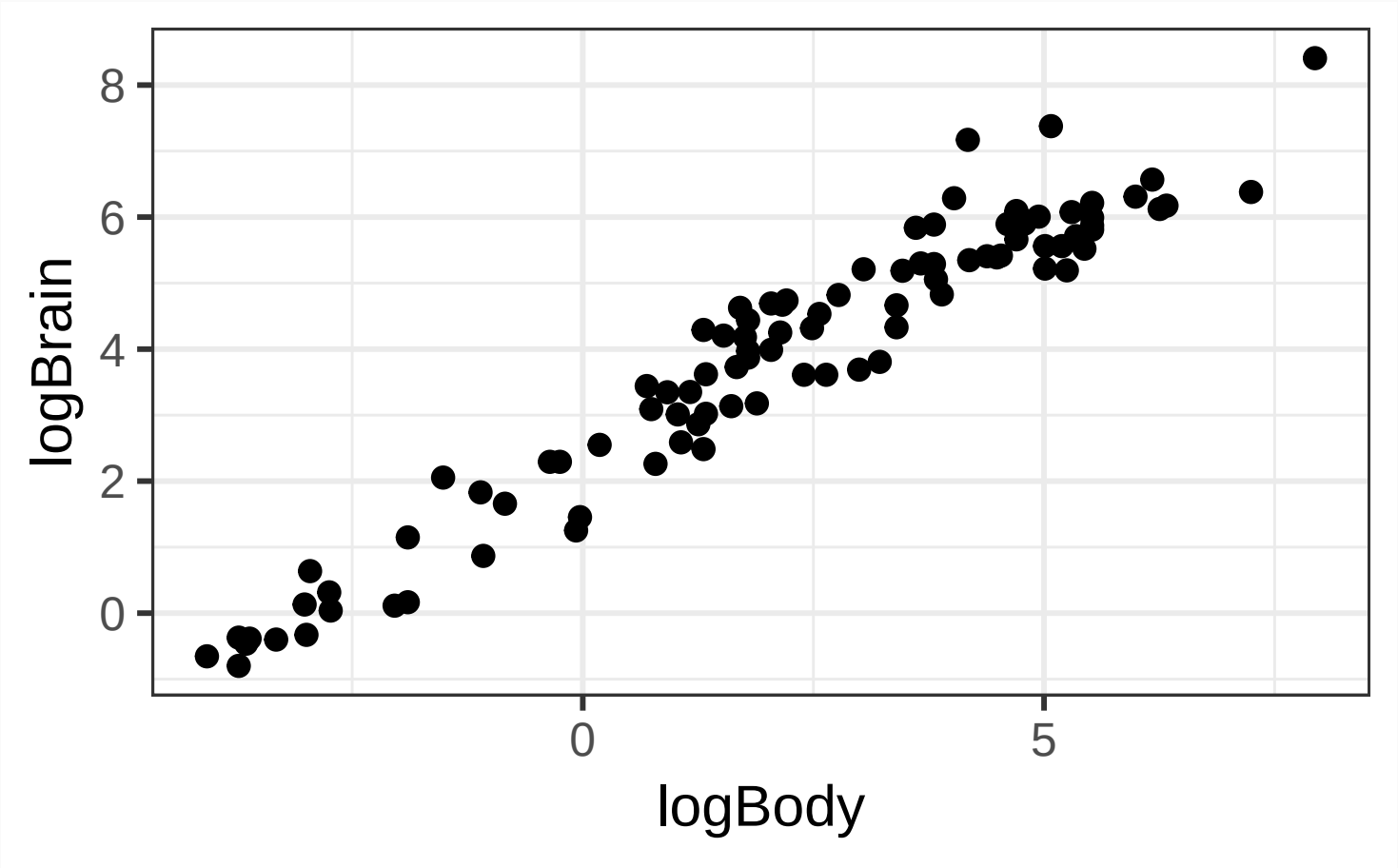


Response against each explanatory variable

- Still a lot of curvature, so it looks like logging each variable might help.

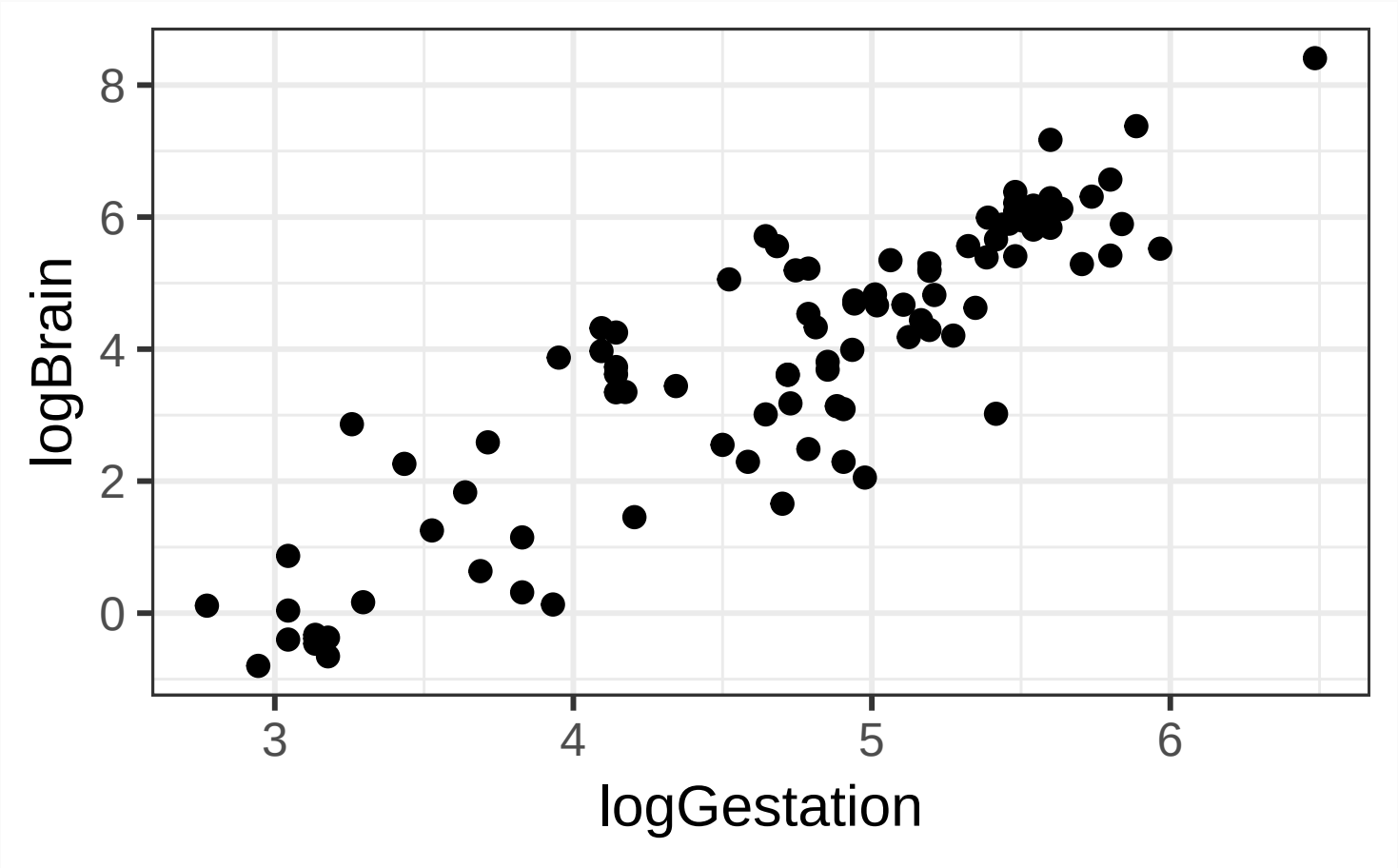
Response against each explanatory variable

```
qplot(logBody, logBrain, data = case0902)
```



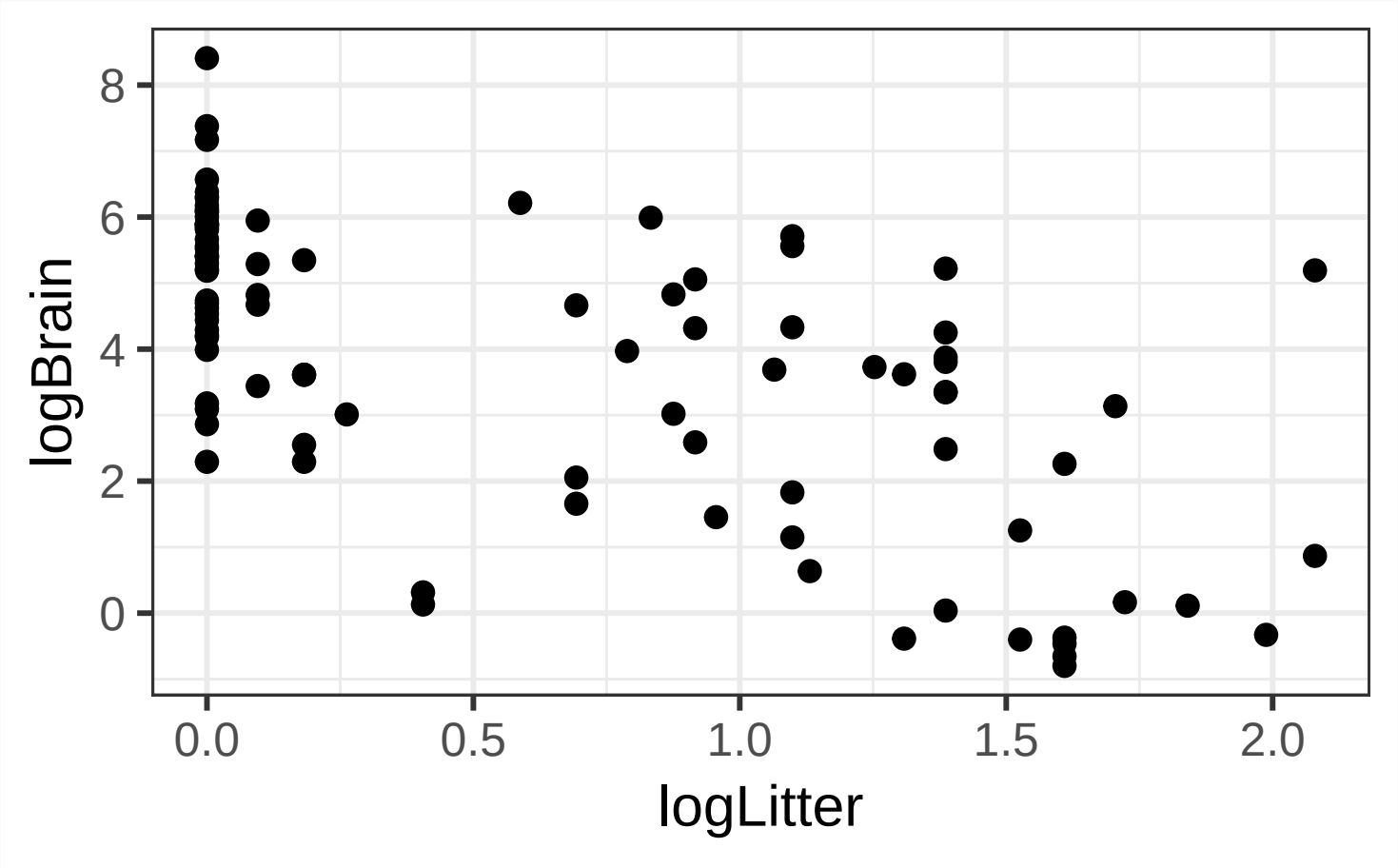
Response against each explanatory variable

```
qplot(logGestation, logBrain, data = case0902)
```



Response against each explanatory variable

```
qplot(logLitter, logBrain, data = case0902)
```



Response against each explanatory variable

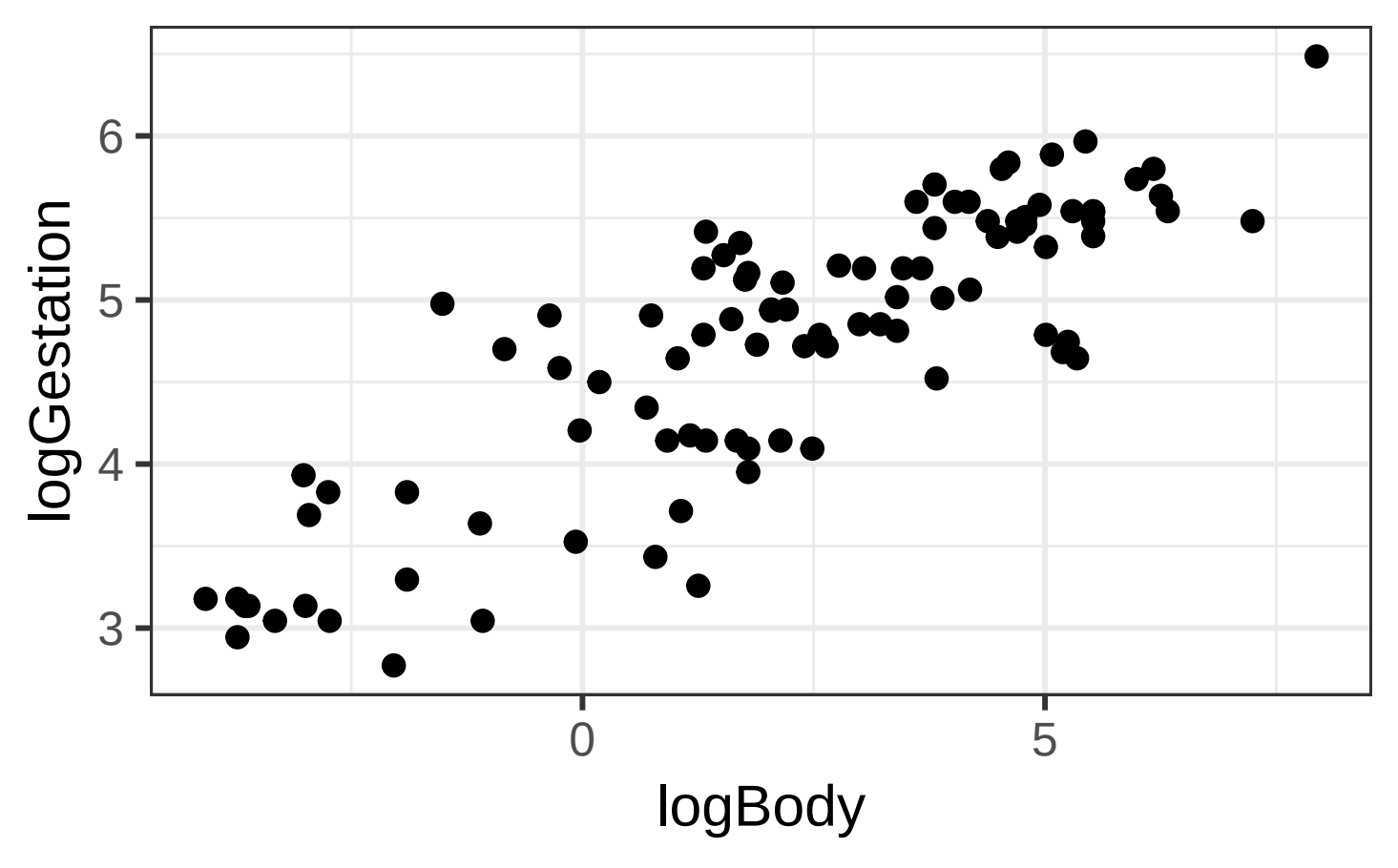
- There looks like there is a lot of linearity now.

Explanatory variables against each other

- It is often useful to look at scatterplots between each pair of explanatory variables.
- This tells us if some variables seem to be picking up a lot of the same information
- E.g. Gestation period might be larger just because body size is larger.
    - If brain is associated with Gestation, it might only be through body.

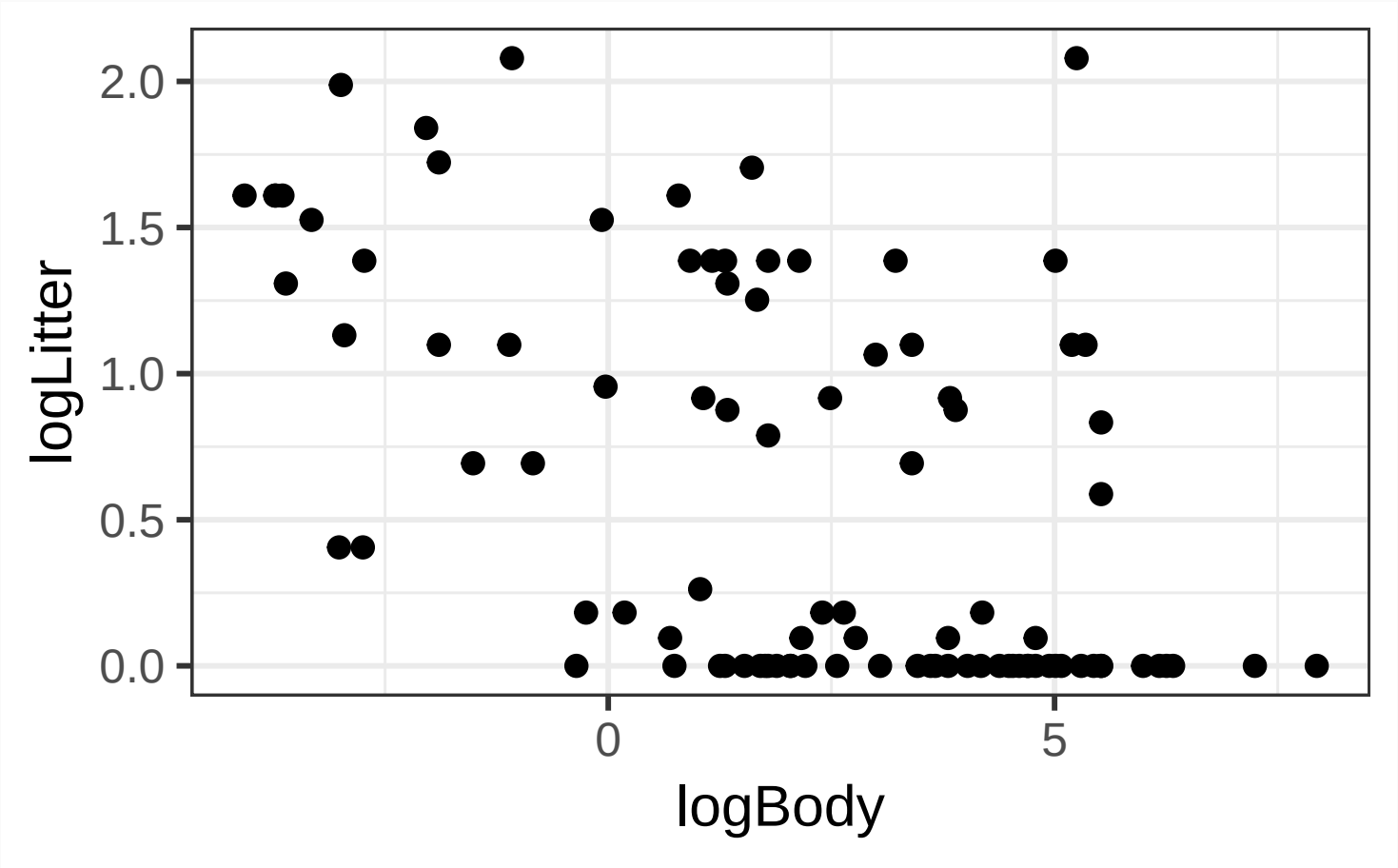
Explanatory variables against each other

```
qplot(logBody, logGestation, data = case0902)
```



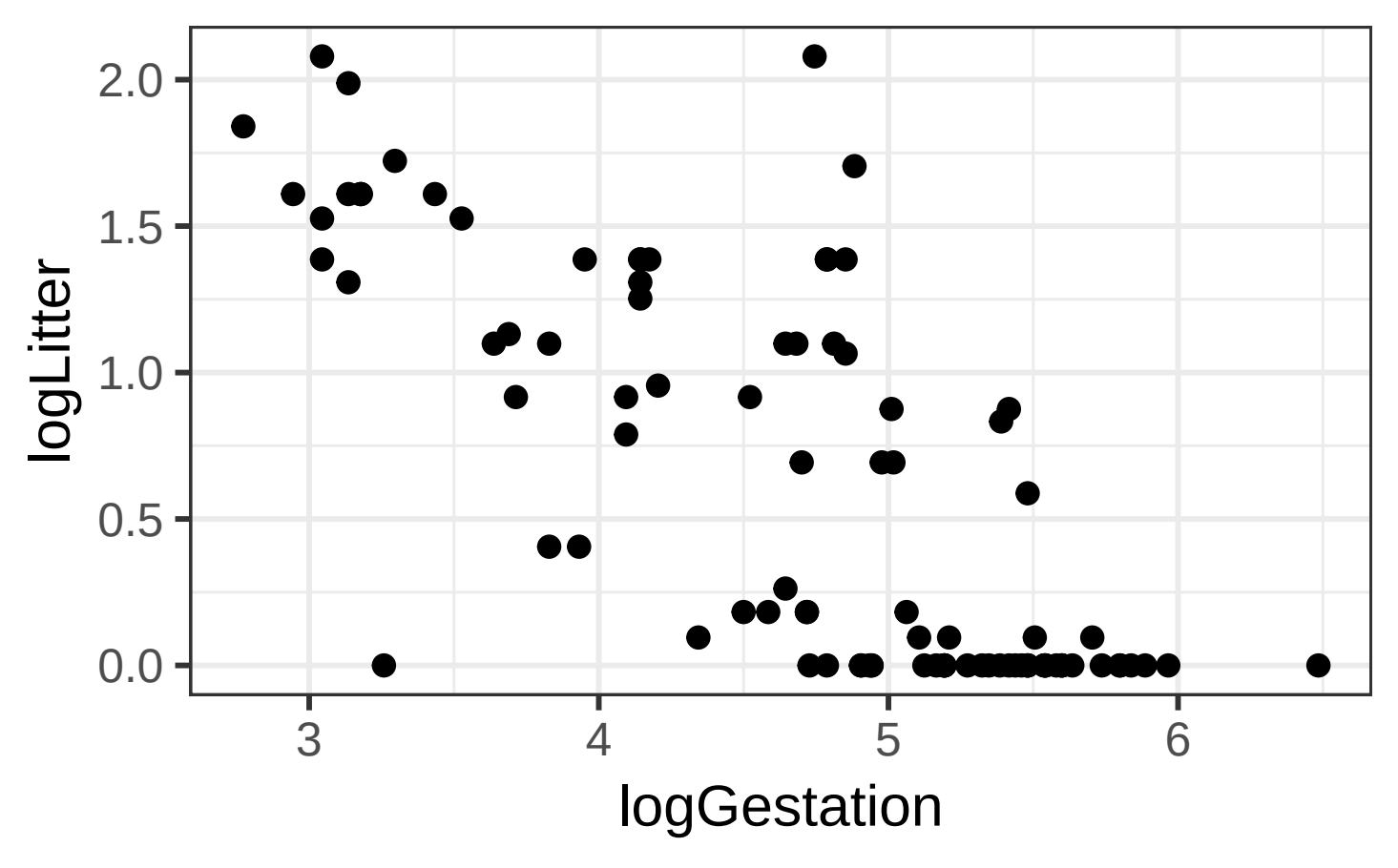
Explanatory variables against each other

```
qplot(logBody, logLitter, data = case0902)
```



Explanatory variables against each other

```
qplot(logGestation, logLitter, data = case0902)
```

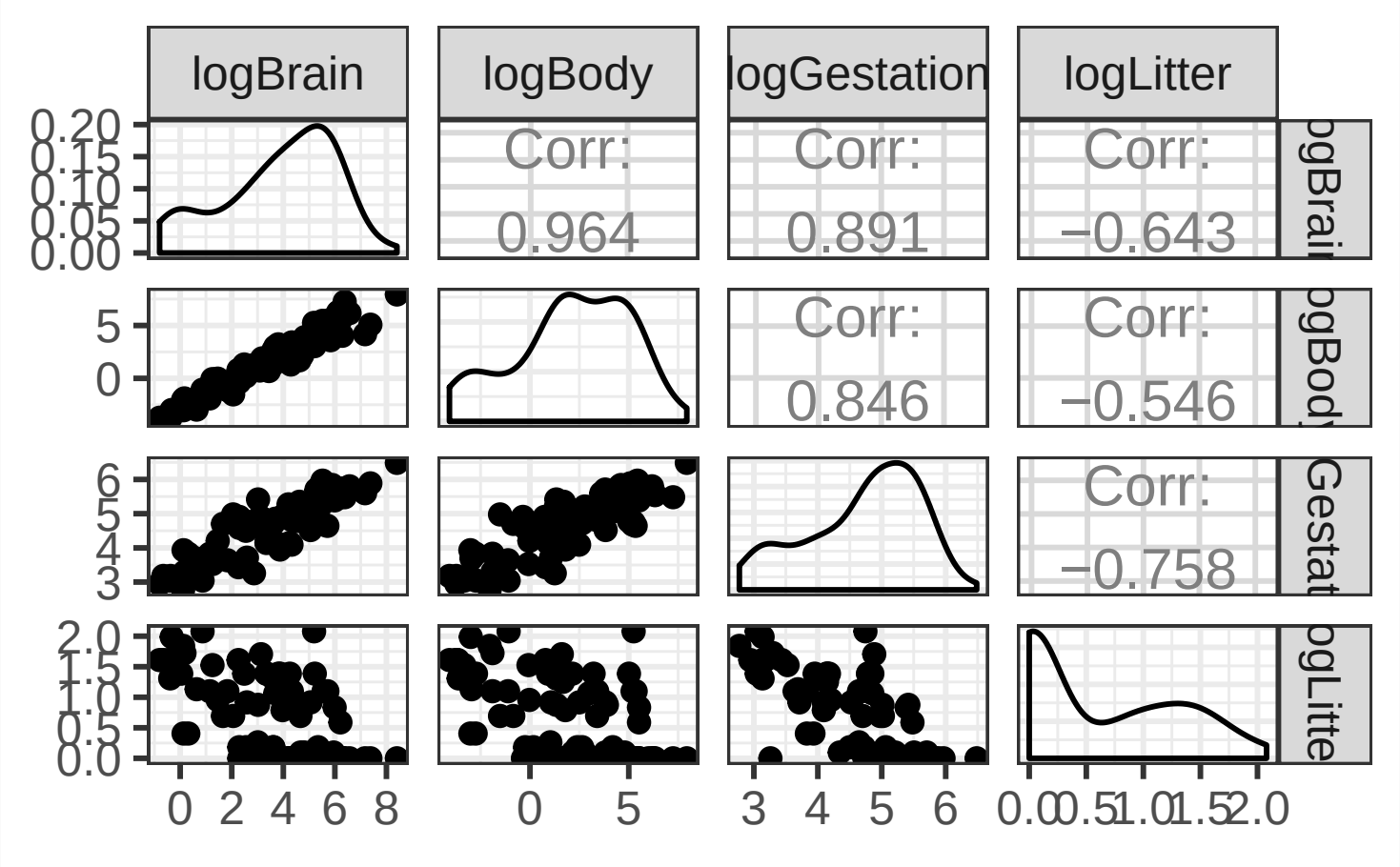


Matrix Plots

- You can show scatterplots between all variables at the same time.
- This is called a matrix plot (or a pairs plot).

Matrix Plots

```
library(GGally)
ggpairs(case0902, columns = 6:9)
```



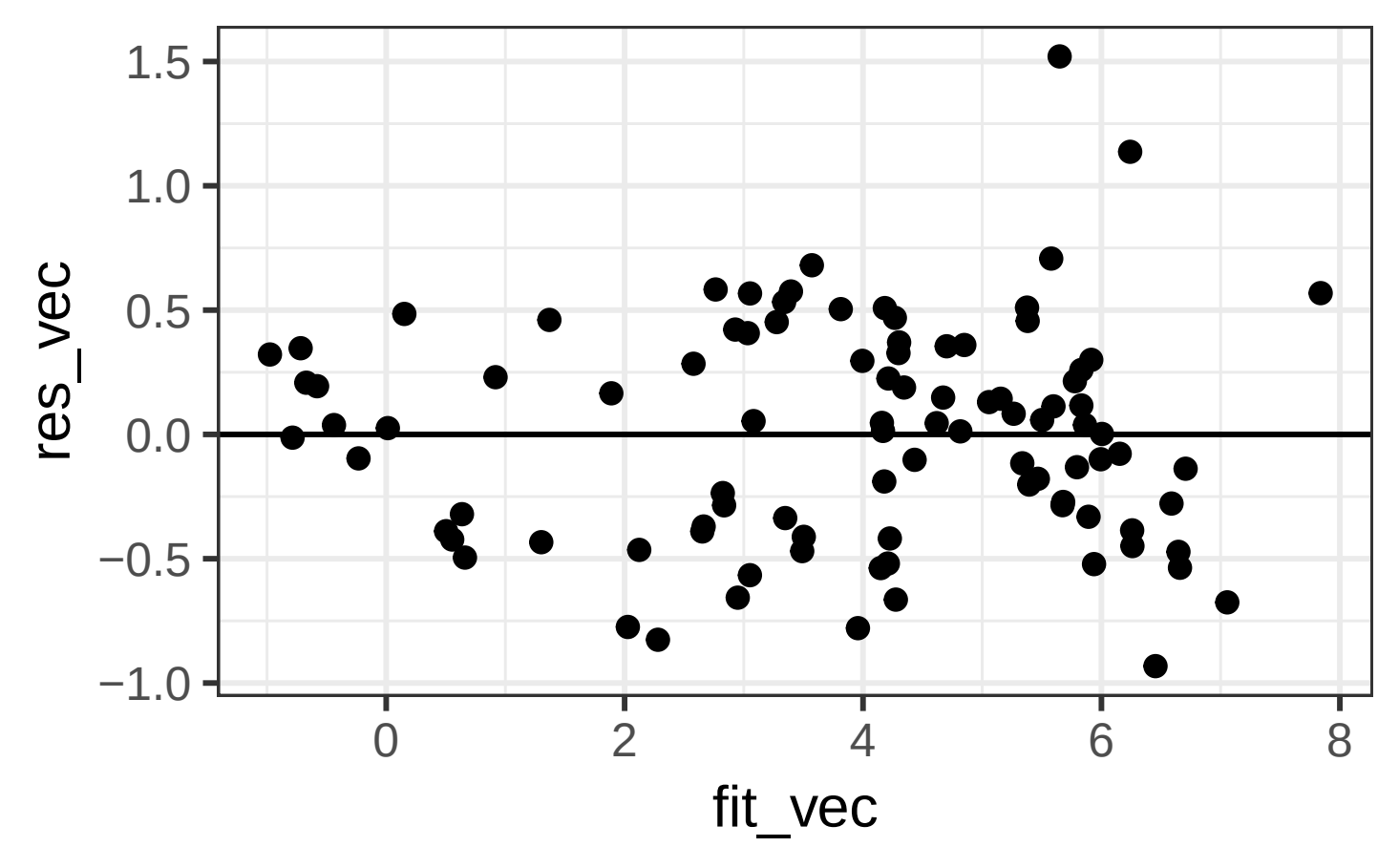
Make an initial fit

- We can first fit a very complicated model and check residuals.
- We would be looking for more curvature and outliers.

```
lm_comp <- lm(logBrain ~ logBody * logLitter * logGestation
              data = case0902)
res_vec <- resid(lm_comp)
fit_vec <- fitted(lm_comp)
```

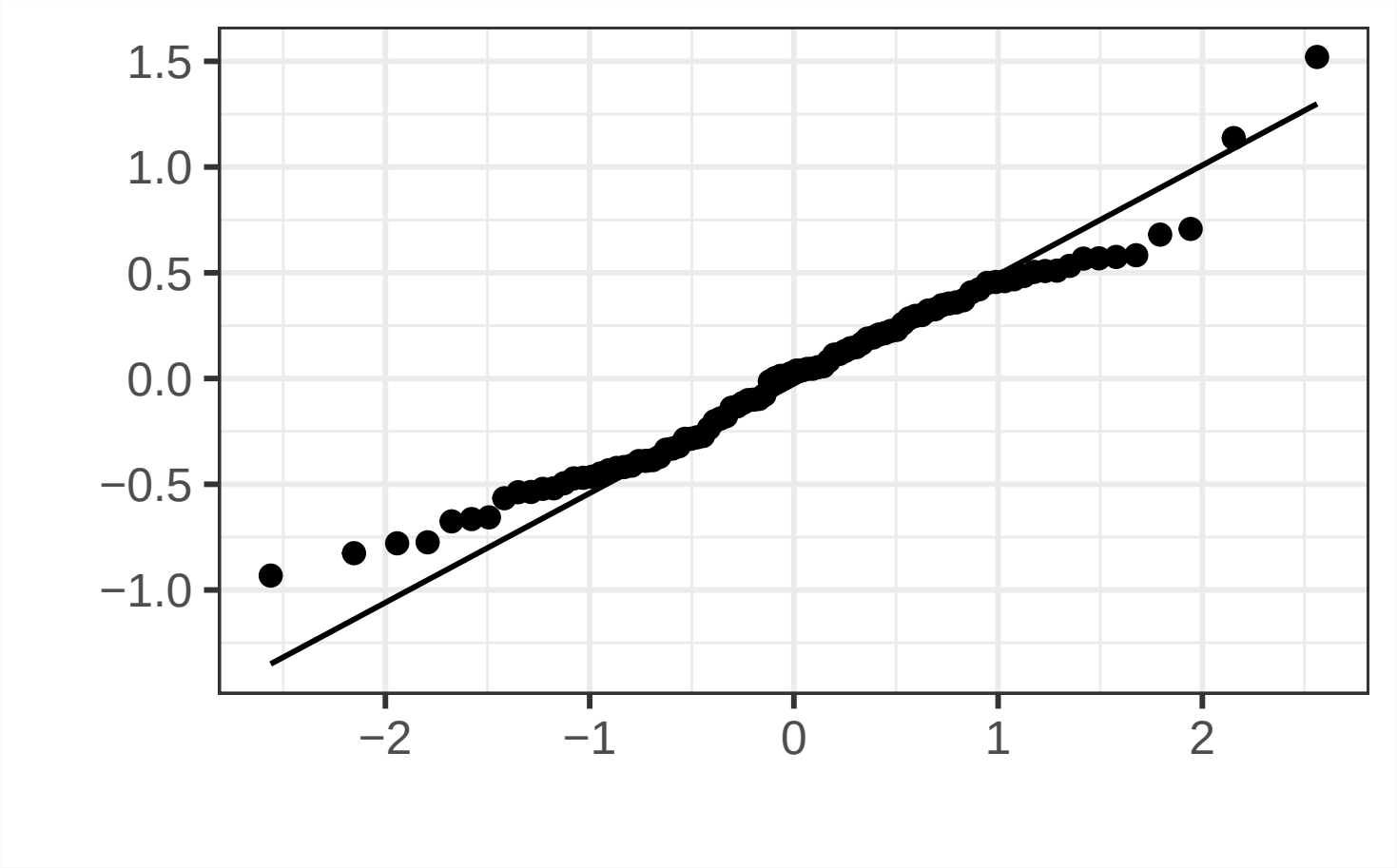
Make an initial fit

```
qplot(fit_vec, res_vec) +
  geom_hline(yintercept = 0)
```



Residual qq-plot

```
qplot(sample = res_vec, geom = "qq") +
  geom_qq_line()
```

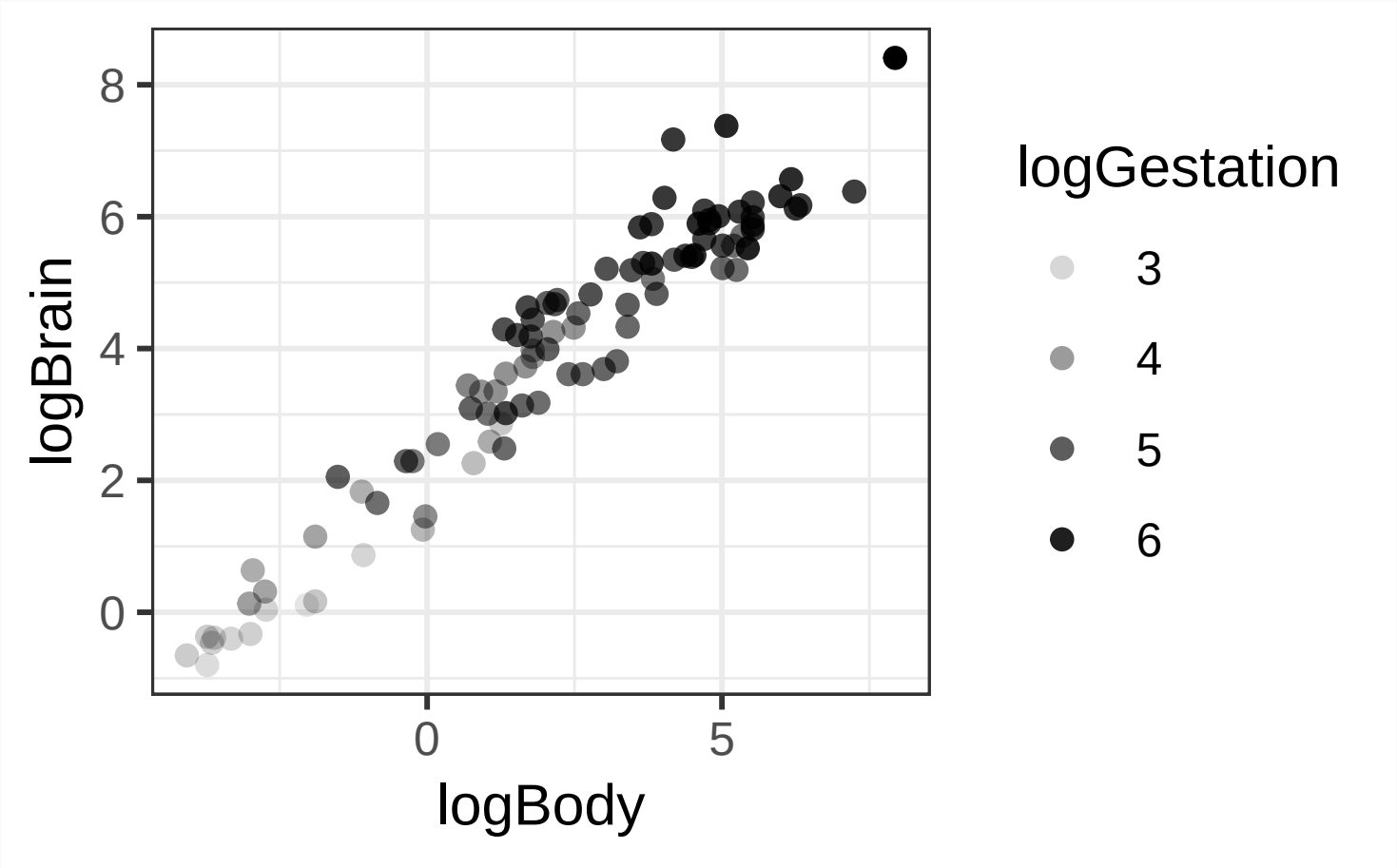


Coded scatterplot

- If you want to explore the association between three quantitative variables, you can code one of them by **transparency** (more preferrable) or **size** (less preferrable)

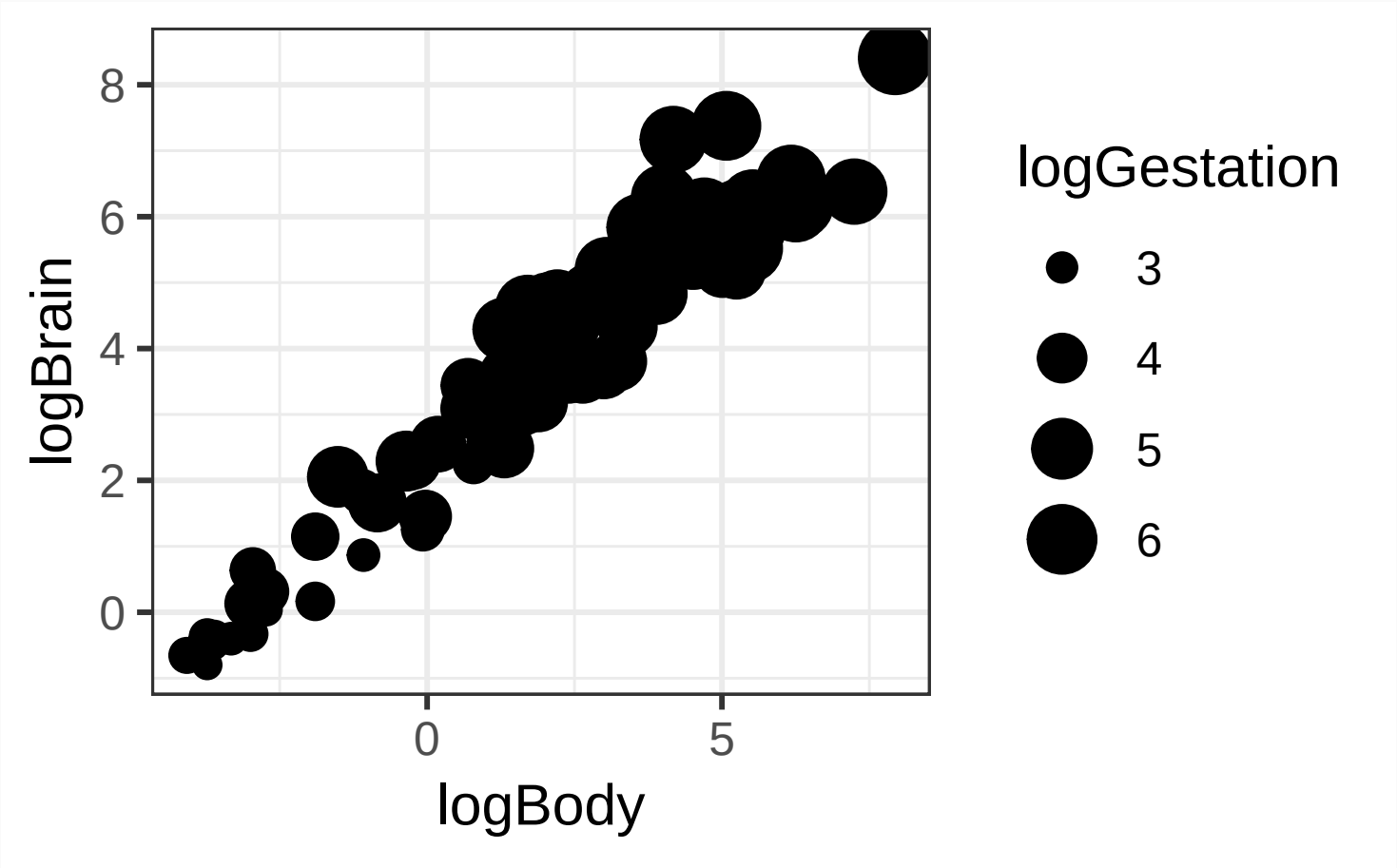
Coded scatterplot

```
qplot(logBody, logBrain, alpha = logGestation,
      data = case0902)
```



Coded Scatterplots

```
qplot(logBody, logBrain, size = logGestation,
      data = case0902)
```

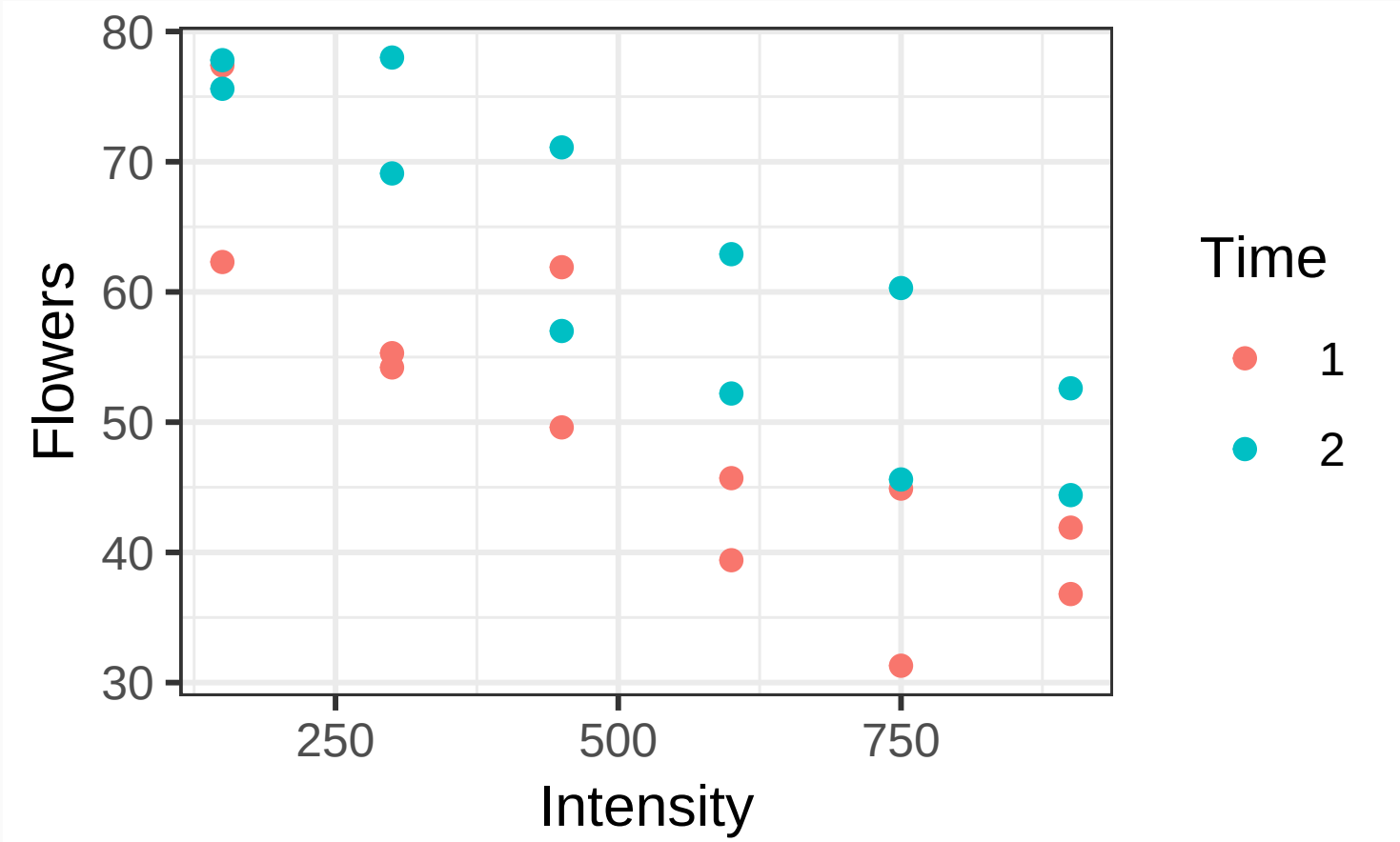


Coded Scatterplots

- If you have categorical explanatory variables, you can code their levels by **colors** (more preferrable) and **shapes** (less preferrable) and include this on a scatterplot of two quantitative variables

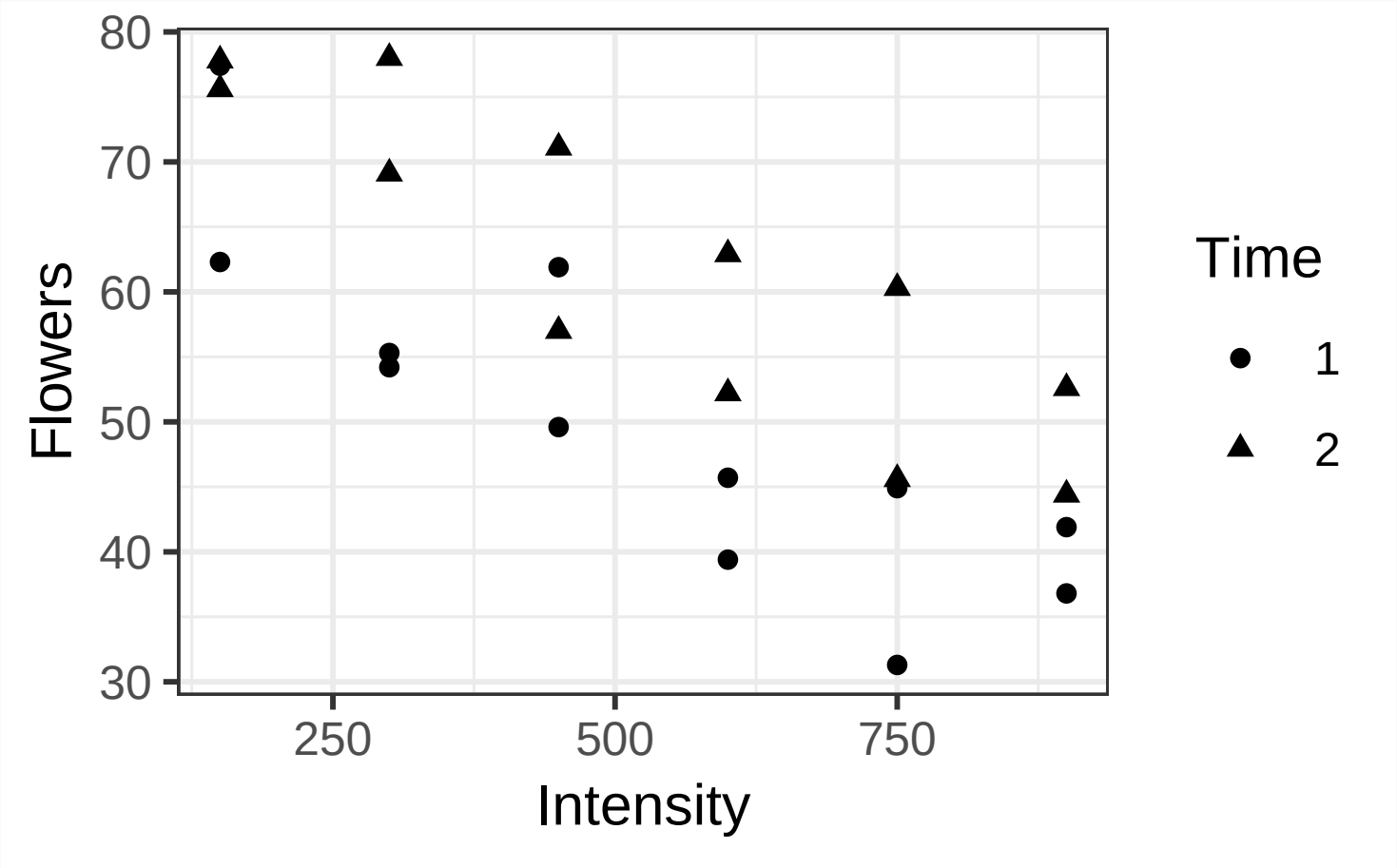
Coded Scatterplots

```
data("case0901")
case0901$Time <- as.factor(case0901$Time)
qplot(Intensity, Flowers, color = Time,
      data = case0901)
```



Coded Scatterplots

```
qplot(Intensity, Flowers, shape = Time,
      data = case0901)
```



Coded Scatterplots

- It's polite to use colorblind safe color palattes

```
library(ggthemes)
qplot(Intensity, Flowers, color = Time,
      data = case0901) +
  scale_color_colorblind()
```