Homework 7

Shun-Lung Chang, Dilip Hiremath

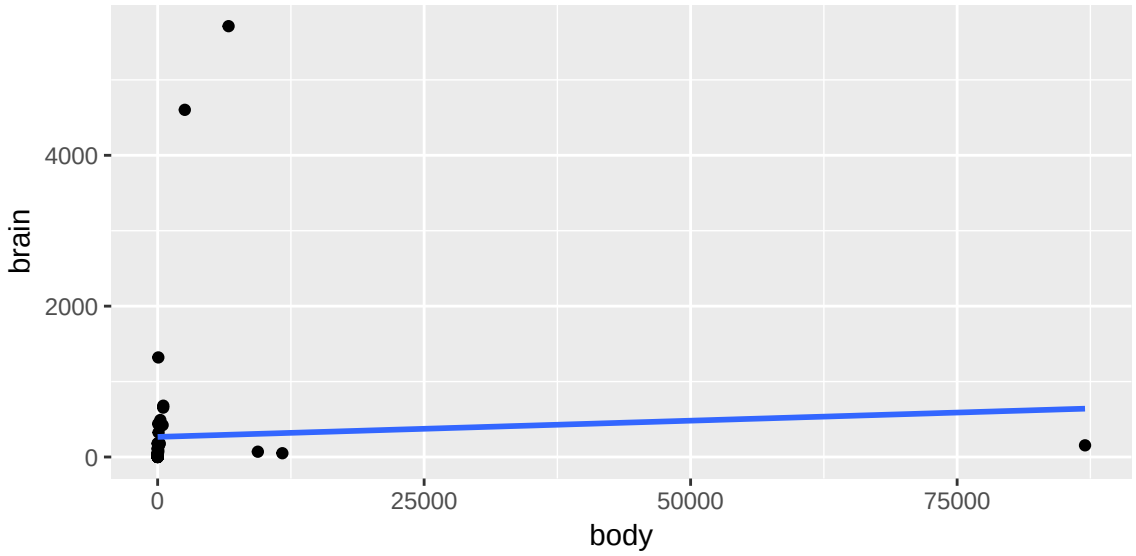
```
load('data/animals.RData')
```

```
library(ggplot2)
```

1. In a first step you graphically check the linearity of the relationship and look for potential transformations to improve linearity.

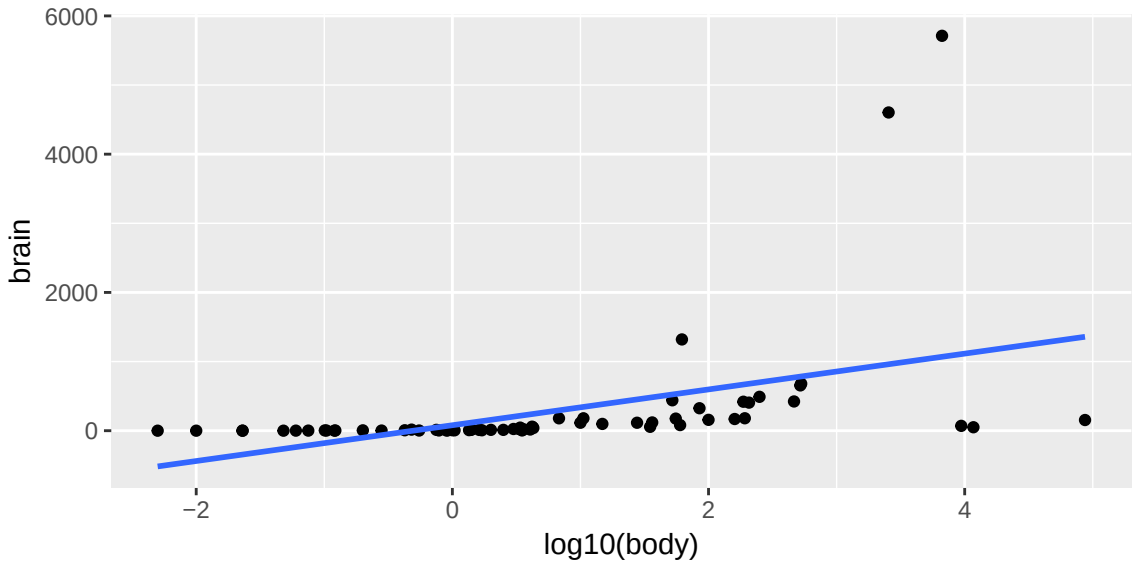
(a) Assess whether there is a linear relationship between brain weight and body weight by looking at a scatterplot of brain weight versus body weight. [Hint: The scatterplot command in Rcmdr provides a ready-made function for adding the regression line and marginal boxplots.]

```
ggplot(data = animals, aes(x = body, y = brain)) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE)
```

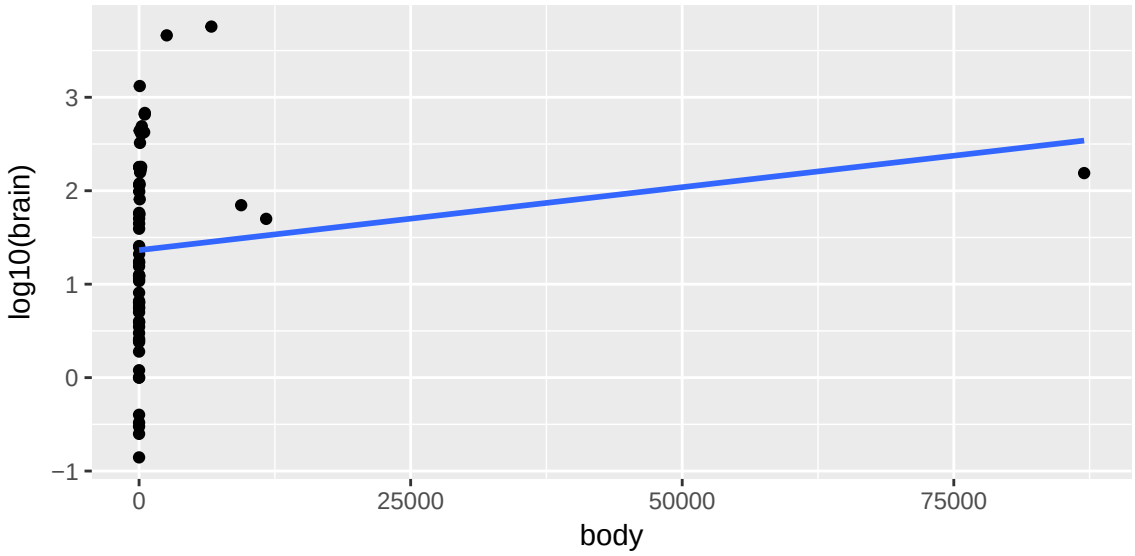


(b) Now, plot the variables on a logarithmic scale (use the logarithm to base 10 here). Draw three scatterplots: one for either of the two scales transformed and one with both scales transformed. Which scatterplot shows the clearest linear relationship?

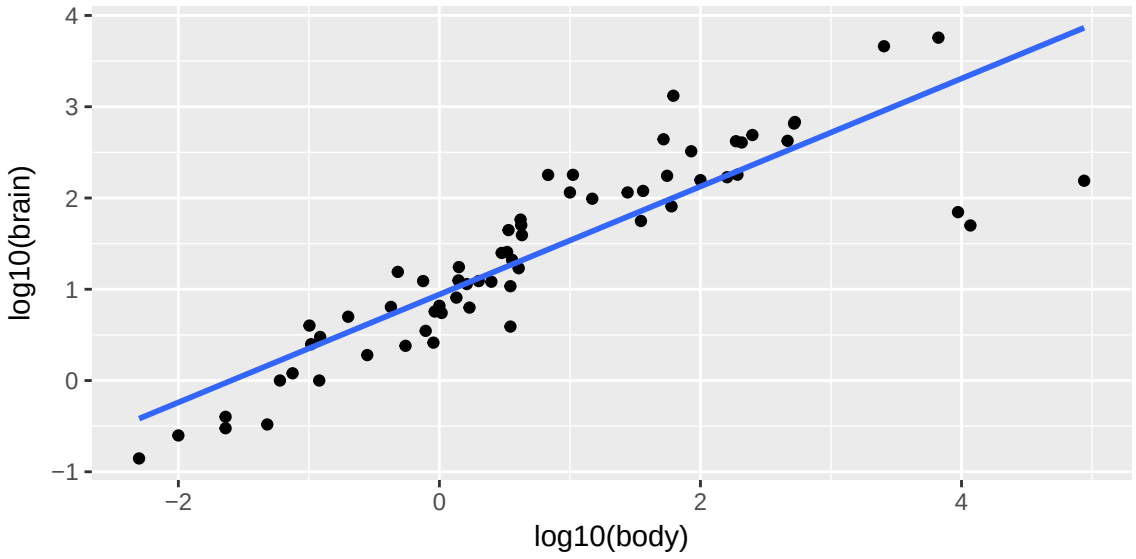
```
ggplot(data = animals, aes(x = log10(body), y = brain)) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE)
```




```
ggplot(data = animals, aes(x = body, y = log10(brain))) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE)
```

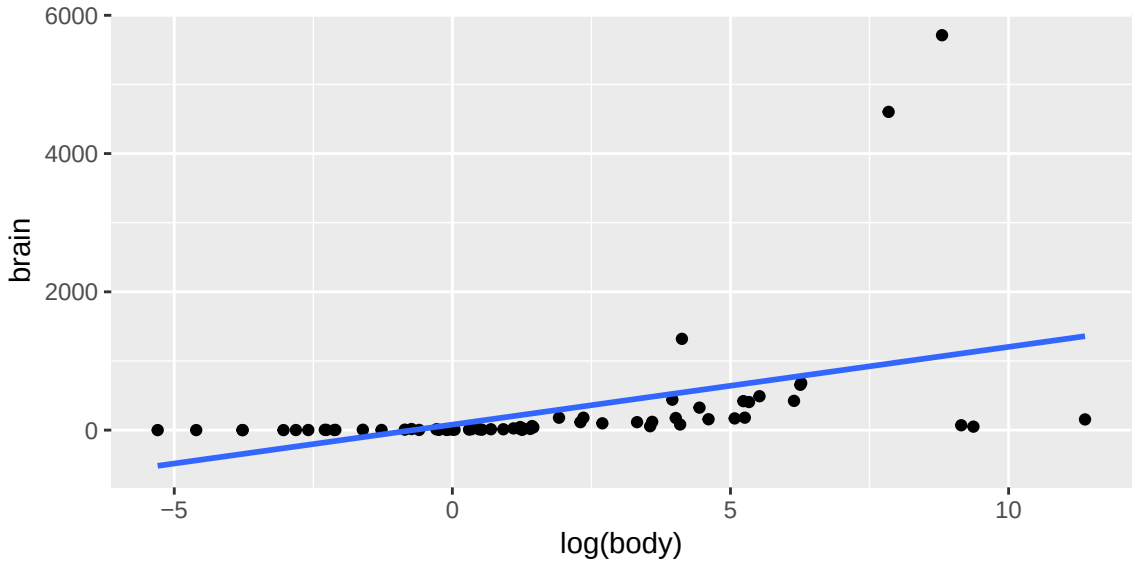



```
ggplot(data = animals, aes(x = log10(body), y = log10(brain))) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE)
```

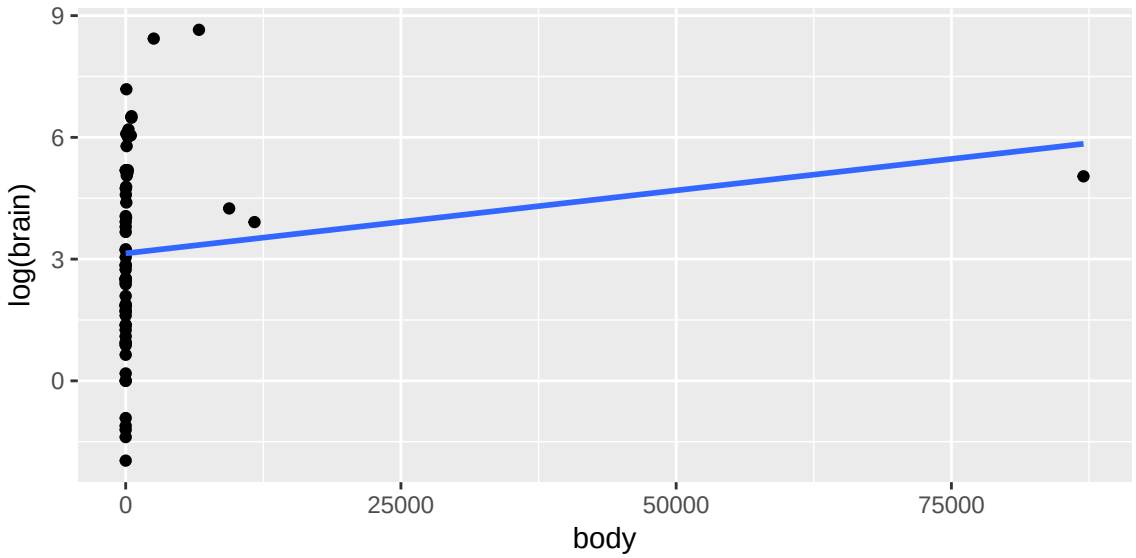


(c) Perform logarithmic transformations (using the natural logarithm with base e) for body weight and brain weight and draw three scatterplots: one for either of the two variables in the original form and the other transformed, and one for both variables transformed. Which scatterplot shows the clearest linear relationship? How do the plots here differ form the ones obtained in Question 1b?

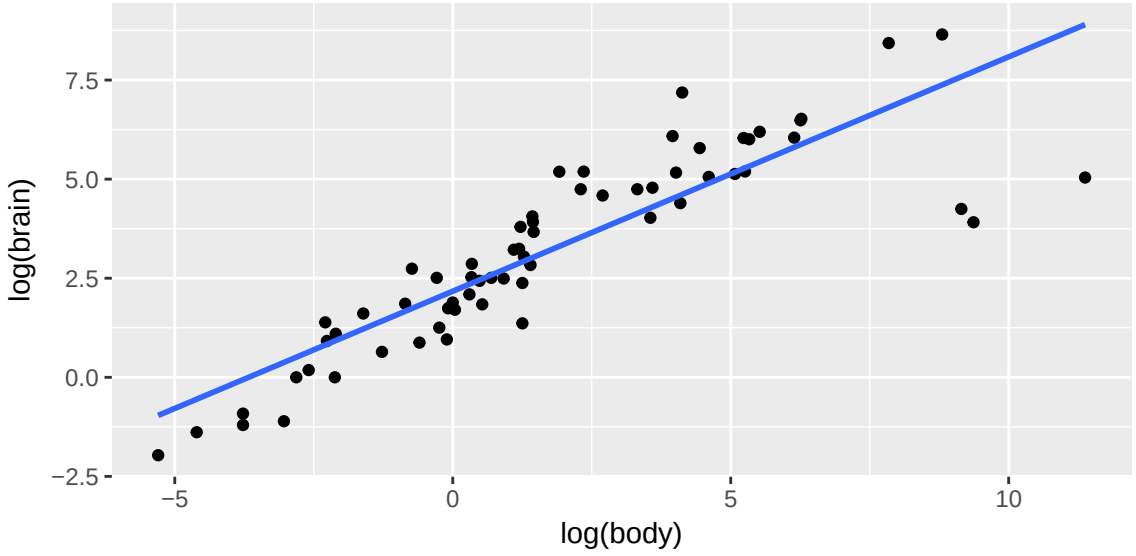
```
ggplot(data = animals, aes(x = log(body), y = brain)) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE)
```




```
ggplot(data = animals, aes(x = body, y = log(brain))) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE)
```




```
ggplot(data = animals, aes(x = log(body), y = log(brain))) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE)
```



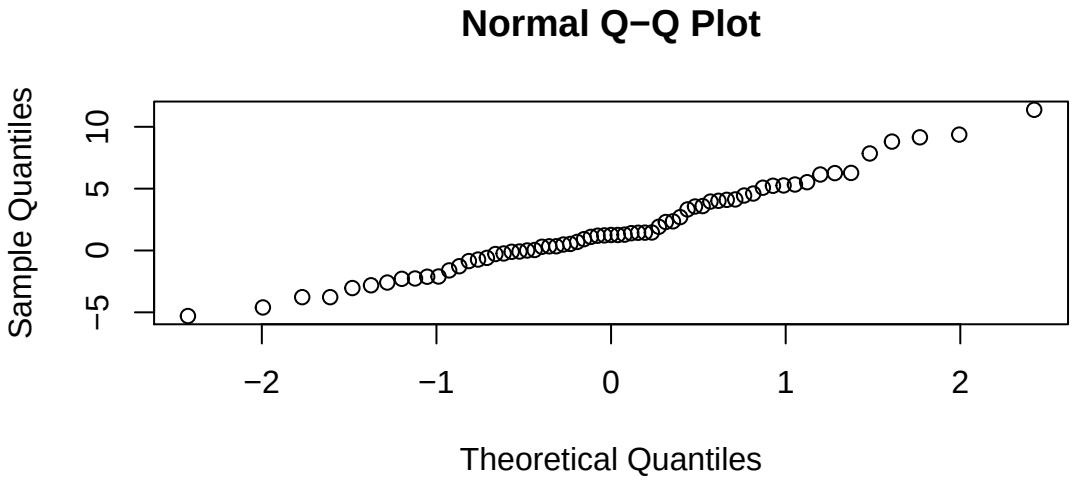
2. In a second step, you explore correlation and linear regression models on this data set and perform some model checks as well.

```
animals$log_body <- log(animals$body)
animals$log_brain <- log(animals$brain)
```

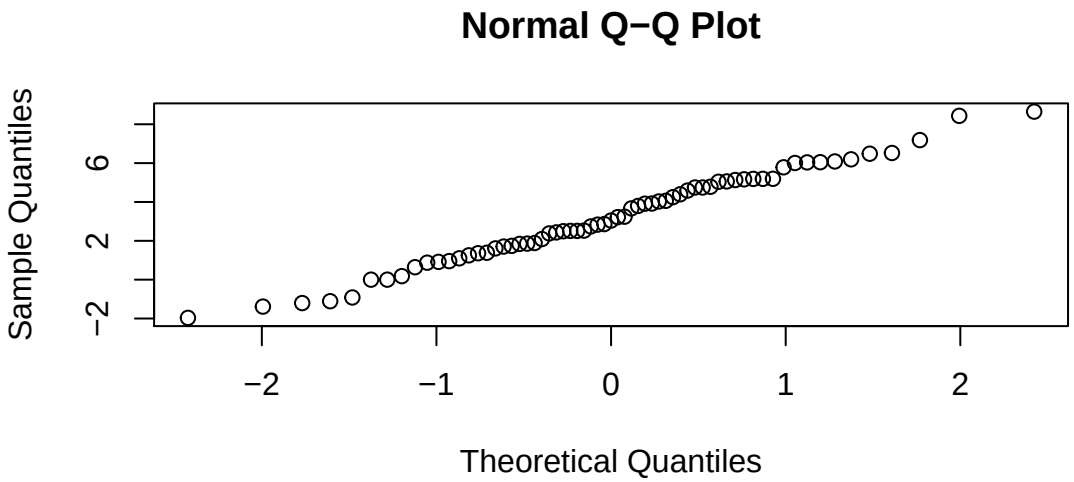
(a) "Homoscedasticity": Graphically inspect whether the variability in scores for logarithmic brain weight is roughly the same at all values of logarithmic body weight.

(b) "Normality": Graphically inspect whether the logarithmically transformed scores for body weight and brain weight are normally distributed. Use the commands qqnorm or qqline for that.

```
qqnorm(animals$log_body)
```




```
qqnorm(animals$log_brain)
```



3. In the next step, you explore correlation and linear regression models on this data set.

(a) Calculate the Pearson correlation coefficient to determine whether logarithmic body weight is related to the logarithmic brain weight. Interpret!

```
cor_coef <- cor(animals$log_body, animals$log_brain)
cor_coef
```

```
[1] 0.8753092
```

(b) Compute a linear regression model for logarithmic body weight depending on the logarithmic brain weight.

```
mod_1 <- lm(log_body ~ log_brain, data = animals)
summary(mod_1)
```

```
Call:
lm(formula = log_body ~ log_brain, data = animals)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7727 -0.8530 -0.1819  0.3027  7.2524

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept) -2.40702    0.35978   -6.69 0.00000000691
log_brain    1.29525    0.09015   14.37       < 2e-16

Residual standard error: 1.735 on 63 degrees of freedom
Multiple R-squared:  0.7662,    Adjusted R-squared:  0.7625
F-statistic: 206.4 on 1 and 63 DF,  p-value: < 2.2e-16
```

4. Compute a linear regression model for logarithmic brain weight depending on the logarithmic body weight. How do you interpret the output in terms of the original variables, body and brain weight?

```
mod_2 <- lm(log_brain ~ log_body, data = animals)
summary(mod_2)
```

```
Call:
lm(formula = log_brain ~ log_body, data = animals)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8592 -0.5075  0.1550  0.6410  2.5724

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.17169    0.16203   13.40   <2e-16
log_body     0.59152    0.04117   14.37   <2e-16

Residual standard error: 1.172 on 63 degrees of freedom
Multiple R-squared:  0.7662,    Adjusted R-squared:  0.7625
F-statistic: 206.4 on 1 and 63 DF,  p-value: < 2.2e-16
```

5. Calculate the standard deviations of logarithmic body and logarithmic brain weights. Then check that the regression slopes obtained in the two models above satisfy the equation.

$bx, y = r\frac{sy}{sx}$

$by, x = r\frac{sx}{sy}$

```
sd_log_body <- sd(animals$log_body)
sd_log_brain <- sd(animals$log_brain)
```

```
slope_1 <- cor_coef * (sd_log_body / sd_log_brain)
slope_1
```

```
[1] 1.29525
```

```
slope_2 <- cor_coef * (sd_log_brain / sd_log_body)
slope_2
```

```
[1] 0.5915198
```

6. In the scatter plot using logarithmically transformed brain and body weight, you can see three observations on the very right of the plot representing animals having rather large values for body weight and respectively small values for brain weight. Which animals are these? Compute a linear regression model that leaves out these points. Did the quality of the model as measured by adjusted R-squared improve? Why?

```
animals[animals$log_body > 9, 'name']
```

```
[1] Dipliodocus
[2] Triceratops
[3] Brachiosaurus
65 Levels: African elephant ...
```

```
# remove outliers
animals_new <- animals[animals$log_body < 9, ]

mod_3 <- lm(log_brain ~ log_body, animals_new)
summary(mod_3)
```

```
Call:
lm(formula = log_brain ~ log_body, data = animals_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43597  1.94829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16
log_body     0.75169    0.02846   26.41   <2e-16

Residual standard error: 0.6943 on 60 degrees of freedom
```

```
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

7. Does the regression model in Question 6 prove that a higher body weight causes a higher brain weight?

Yes, given the positive coefficient, we can conclude that a higher body weight results in a higher brain weight.

8. Now you are using the model obtained in Question 6 to predict brain weight for some animals.

(a) Which brain weight would you predict for a Southern long-nosed armadillo with a body weight of 3.6 kg, and which for a female blue whale with a body weight of 150 tons?

```r
predict(mod_3, data.frame(log_body = log(3.6)))
```

```
       1
3.097649
```

```r
# 150 tons is around 136078kg
predict(mod_3, data.frame(log_body = log(136078)))
```

```
       1
11.02046
```

(b) Which one of the two predictions you just made do you find more reliable? Why?

The first prediction is more reliable, since the predictions cannot be safely extrapolated to values beyond the maximal body weight in our dataset (87000).