Linear Regression

Andre Bourbonnais

2023-11-20

Linear regression

```
library(tidyverse)
library(ggpubr)
```

The linear model:

$$y_i = \beta_0 + \sum x_{ij}\beta_j x_i + \epsilon_i$$

The tern β_0 is the *intercept* (also known as α), which in the context of a linear regression gives the value of the *response variable* y when the *predictor variable* x is equal to zero. The β_j are the *coefficients*, aka the the *slopes* for the predictor variables x, and the ϵ_i represents the *residuals*. The residuals are the deviations of each data point from its expected value based on the *fitted model*. The *linear model* assumes that the residuals are normally distributed.

The aim of regression

The aim of regression is to estimate the linear *relationship* between the response variable y (dependent variable) and one or more predictor variables (x), also known as *independent variables*. The linear relationship is defined by the *coefficients* β_j. The coefficients are estimated by minimizing the sum of the squared residuals, also known as the *least squares method*. Where the regression parameters are estimated by minimizing the sum of the squared residuals to best fit the estimated regression line.

Linear regression basics

```
# Simulate data
set.seed(1337)
x <- rnorm(n=200, mean = 10, sd = 2)
y = 0.4*x + rnorm(n=200, mean = 0, sd = 1)

# Linear model
m <- lm(y~x)

# The coefficients
m$coef
```

```
## (Intercept)           x
##   0.1344746   0.3773854
```

The object `m` contains the results of the linear model where the *coefficients* are stored in the `m$coef` object. The first value is the intercept, and the second value is the slope. The **intercept** is the value of the **response variable** when the **predictor variable is equal to zero**. The slope is the change in the response variable for a **one unit change in the predictor variable.**

Meaning, the slope intercepts the *y-axis* at 0.134, and for each x value increase, the y value increases by 0.377. As can be seen in the plot below where **A** shows the regression line and **B** shows the intercept.

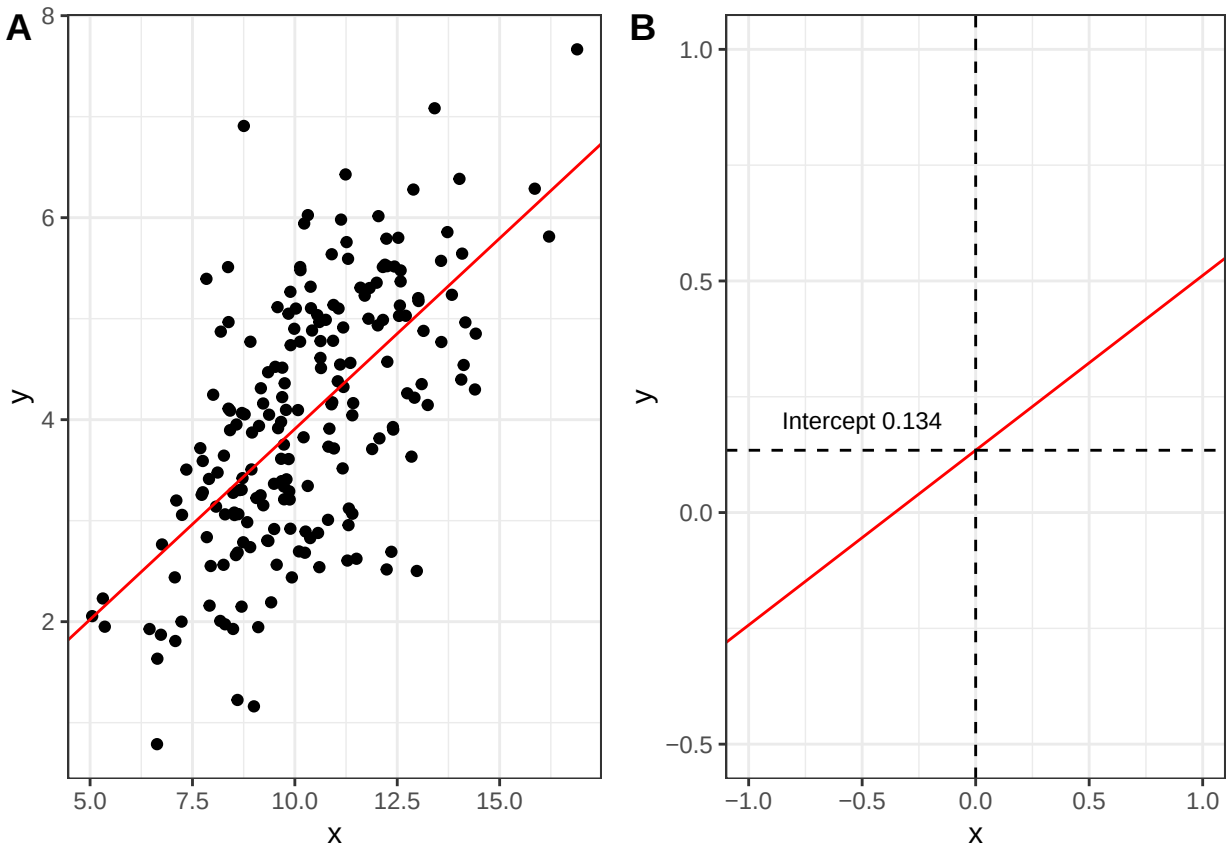
```
# Plot the data and the regression line using m$coef with ggplot.
p1 <- ggplot() +
  geom_point(aes(x = x, y = y)) +
  geom_abline(intercept = m$coef[1], slope = m$coef[2], color = "red") +
  theme_bw()

p2 <- ggplot() +
  geom_point(aes(x = x, y = y)) +
  geom_abline(intercept = m$coef[1], slope = m$coef[2], color = "red") +
  xlim(-1, 1) +
  ylim(-0.5, 1) +
  geom_vline(xintercept = 0, color = "black", linetype = "dashed", size = 0.5) +
  geom_hline(yintercept = m$coef[1], color = "black", linetype = "dashed",
             size = 0.5) +
  annotate("text", x = -0.5, y = 0.2, label = "Intercept 0.134", size = 3) +
  theme_bw()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
ggarrange(p1, p2, ncol = 2, nrow = 1, labels = c("A", "B"))
```

```
## Warning: Removed 200 rows containing missing values (‘geom_point()‘).
```



Understanding the `summary()` The summary function contains a lot of information.

```
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5304 -0.6994  0.0452  0.6913  3.4699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.13447    0.35472   0.379    0.705
## x            0.37739    0.03383  11.155   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9912 on 198 degrees of freedom
## Multiple R-squared:  0.3859, Adjusted R-squared:  0.3828
## F-statistic: 124.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

Residual distribution The distribution of the residuals can be found in the **Residuals:** part of the summary output. We get the quantiles and the median. As the 1st and 3rd quantile is symmetrical we can

quickly see that the residuals are normally distributed.

Model parameters (Coefficients) The **Coefficients:** part of the summary output contains the estimated coefficients, the standard error, the t-value, and the p-value. The estimated coefficients are the same as the `m$coef` object.

Slope Although the regression slope is often reported without any units, it is important to remember that the slopes in fact carry the units of both the response and the predictor variables. Imagine that the x and y are measured in mm, the slope is 0.377, which means that for each 1 mm increase in x, y increases by 0.377 mm. Meaning, 0.377mm/mm. When reporting the slope we also want to report the *Standard Error* of the slope. The standard error is a measure of the uncertainty of the slope estimate and can be fetched in `R` by `summary(m)$coef[2,2]`.

```
summary(m)$coef[2,2]
```

```
## [1] 0.03383184
```

Thus, we should report the slope as $0.377 \pm 0.034\, mm/mm$. The small standard error (relative to the slope estimate) indicates that the slope is estimated with high precision.

Futhermore, say we want to know how much y increases for one *standard deviation* increase in x. To do this we compute the difference between $f(mean(x) + sd(x)) - f(mean(x))$. Therefore we can say, *"That y increased by 2.07mm per standard deviation increase in x."*

```
(m$coef[2] * mean(x) + sd(x)) - (m$coef[2] * mean(x))
```

```
##        x
## 2.076789
```

Coefficient of determination, the r^2 Very important parameter, the r^2 is the proportion of the variance in the response variable that is explained by the model. The r^2 is a measure of the *goodness of fit* of the model. The r^2 is the ratio of the *explained variance* to the *total variance*.

We see in our summary: *Multiple R-squared: 0.3859*, which tells us that x explains 38.59% of the variance in y. It can also simply calculated by `cor(x, y)^2`.

```
cor(x, y)^2
```

```
## [1] 0.3859097
```

The plight of p-values

One should not be hostile against p-values but because the p-value is obtained by comparing the observed test statistic t to its known distribution, and t increases with sample size, it follows that when the sample size increases, anything will at some point be *statistically significant*. Thus, there is a push to move away from p-values and focus more on other statistical measures to evaluating and interpreting the results (effect size, confidence intervals). We need to focus on the interpretation of the parameter estimates, their units, and their consequence withing the context of the analysis/study.

Linear regression example

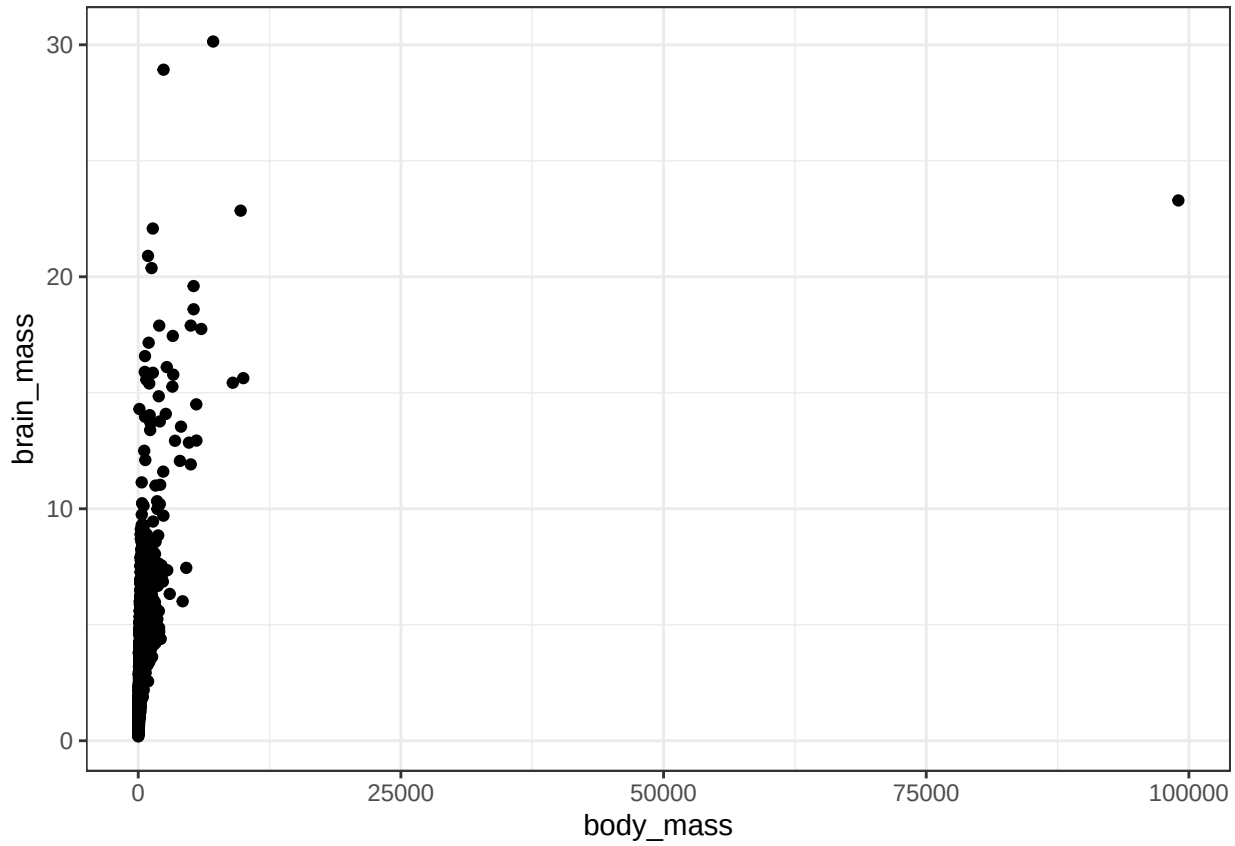
I wonder if there is a difference in allometry between females and birds based on their bird brains yaoo. ## The data

```
raw_df <- read.table(file = "../00_DATA/bird_allometry.csv", header = TRUE,
                     sep = ",")

# Overview of the data
glimpse(raw_df)
```

```
## Rows: 1,184
## Columns: 4
## $ Genus_Species <chr> "Accipiter_gentilis", "Accipiter_gentilis", "Accipiter_n~
## $ Sex           <chr> "f", "m", "f", "m", "f", "m", "m", "m", "f", "m", "f", "~
## $ brain_mass    <dbl> 7.686143, 7.618500, 3.112797, 2.637390, 5.700000, 2.3100~
## $ body_mass     <dbl> 1049.1571, 678.2833, 252.1263, 136.1441, 520.0000, 122.3~
```

```
# Inspection of the data
raw_df %>%
  ggplot(aes(x = body_mass, y = brain_mass)) +
  geom_point() +
  theme_bw()
```



The data analysis

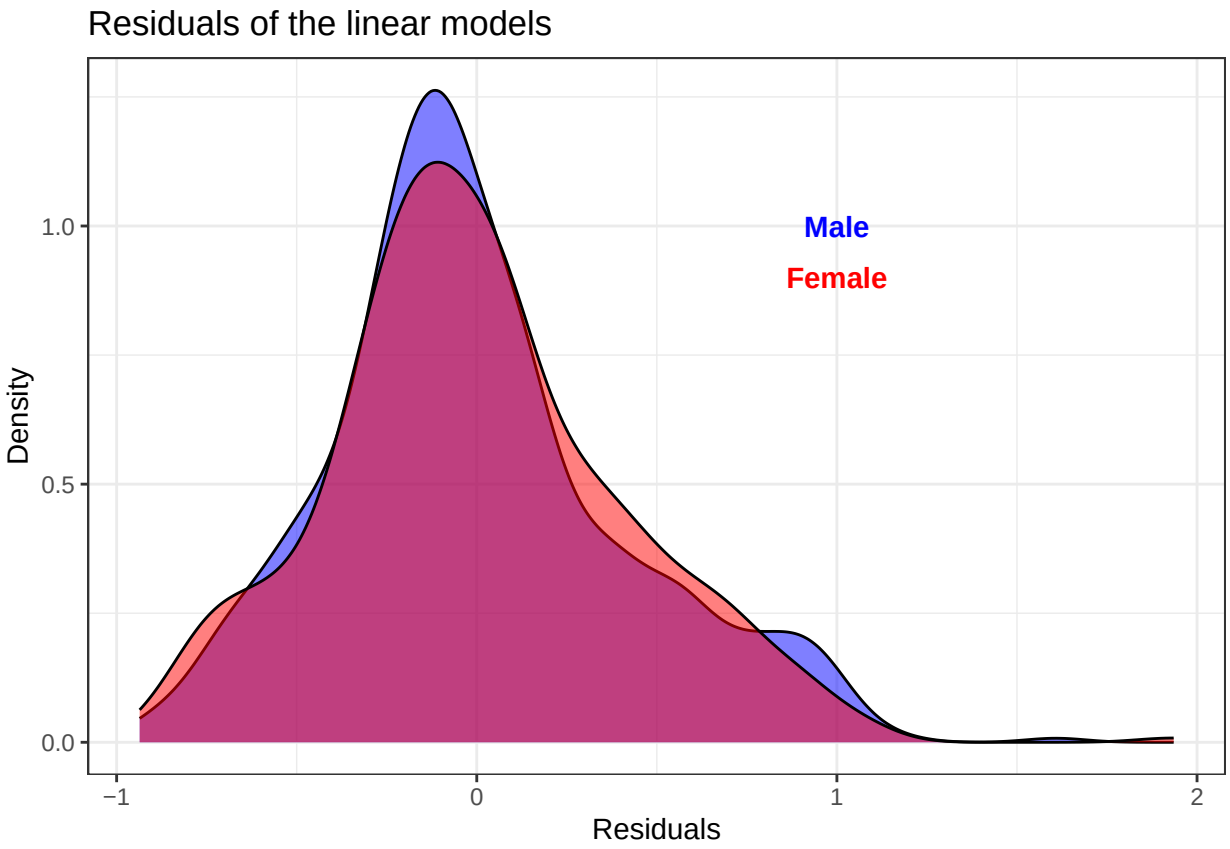
Because the data is expected to follow a *power-law* relationship, we will log-transform the data and then fit a linear model to the data.

```
# We saw that we had one major outlier that we need to deal with.
# So we avoid the outlier by filtering out the body_mass > 25000
# and then remove the NA's if any
df <- raw_df %>%
  filter(body_mass < 25000) %>%
  na.omit()

# Lets fit some models
mm <- df %>%
  select(Sex, body_mass, brain_mass) %>%
  filter(Sex == "m") %>% # Only males
  lm(log(brain_mass) ~ log(body_mass), data = .)

mf <- df %>%
  select(Sex, body_mass, brain_mass) %>%
  filter(Sex == "f") %>% # Only females
  lm(log(brain_mass) ~ log(body_mass), data = .)

# Plot the data to see if they are normaly distributed
ggplot() +
  geom_density(aes(x = residuals(mm)), fill = "blue", alpha = 0.5) +
  geom_density(aes(x = residuals(mf)), fill = "red", alpha = 0.5) +
  labs(x = "Residuals", y = "Density") +
  ggtitle("Residuals of the linear models") +
  theme_bw() +
  annotate("text", x = 1, y = 1, label = "Male", color = "blue",
           fontface = "bold") +
  annotate("text", x = 1, y = 0.9, label = "Female", color = "red",
           fontface = "bold")
```




```
# The residuals looks normaly distributed, as the quantiles are symmetric.
# summary(mm)
# summary(mf)
```

Lets peak some more in the `summary()` of the models. What we can see is that we have an negative intercept which can be explained as we have different species of birds. Lets focus on the slope. We see that there is an ~0.5 increase per 1 increase in body mass. However, we log transformed the data, and thus this mean that the brain size increases by 5.5% per 10% increase in body size. The standard errors are very small which means they are estimated with high precision.

```
print("Male summary")
```

```
## [1] "Male summary"
```

```
summary(mm)
```

```
##
## Call:
## lm(formula = log(brain_mass) ~ log(body_mass), data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93609 -0.25102 -0.06132  0.19585  1.60829
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.926666   0.045658  -42.20   <2e-16 ***
## log(body_mass)  0.555635   0.009977   55.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4114 on 612 degrees of freedom
## Multiple R-squared:  0.8352, Adjusted R-squared:  0.8349
## F-statistic:  3102 on 1 and 612 DF,  p-value: < 2.2e-16
```

```
print("Female summary")
```

```
## [1] "Female summary"
```

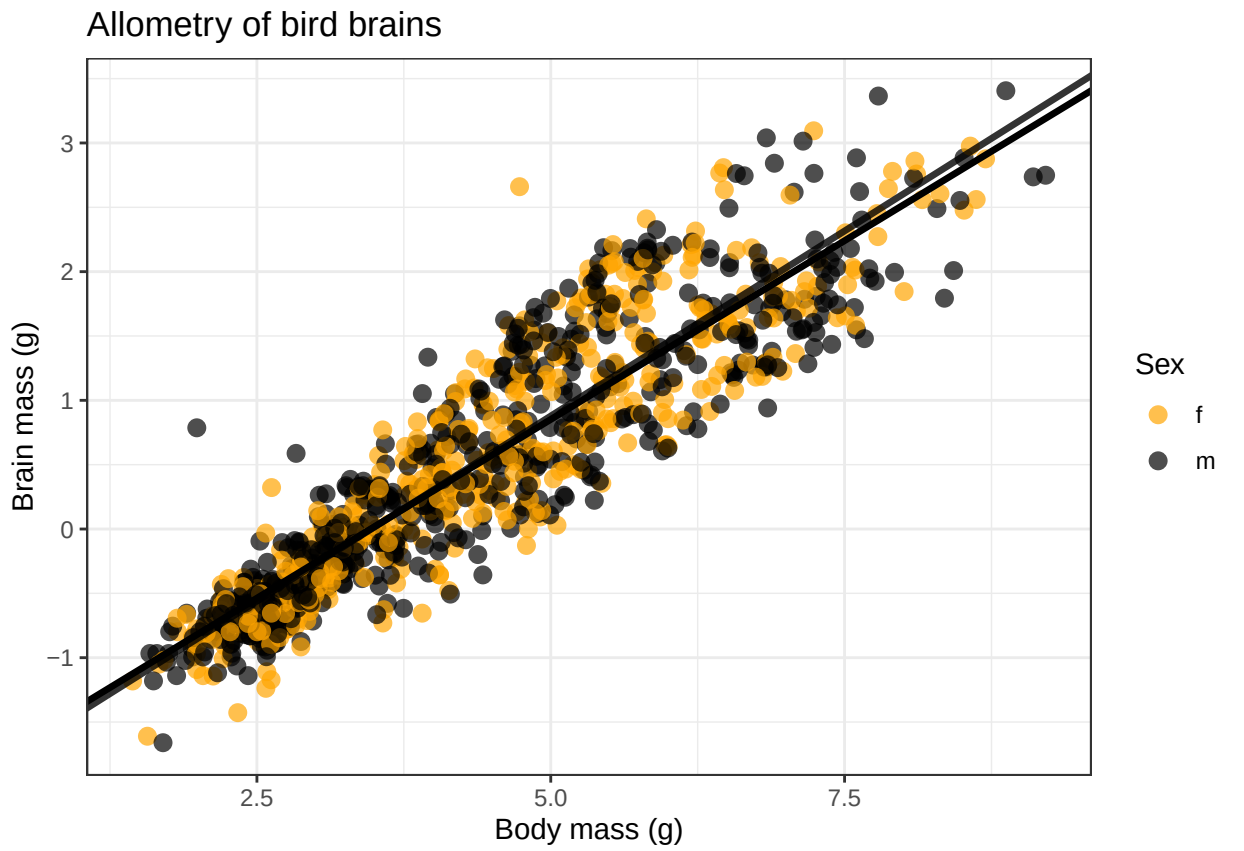
```
summary(mf)
```

```
##
## Call:
## lm(formula = log(brain_mass) ~ log(body_mass), data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90255 -0.26201 -0.03523  0.24946  1.93514
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.99844    0.05532  -36.12   <2e-16 ***
## log(body_mass)  0.57529    0.01184   48.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4116 on 481 degrees of freedom
## Multiple R-squared:  0.8307, Adjusted R-squared:  0.8303
## F-statistic:  2359 on 1 and 481 DF,  p-value: < 2.2e-16
```

Plotting the regression lines

```
#
df %>%
  select(Sex, brain_mass, body_mass) %>%
  filter(Sex != "unknown") %>%
  mutate(log.brain_mass = log(brain_mass),
         log.body_mass = log(body_mass)) %>%
  ggplot(aes(x = log.body_mass, y = log.brain_mass, color = Sex)) +
  geom_point(shape = 16, size = 3, alpha = 0.7) +
  scale_color_manual(values = c("orange", "black")) +
  geom_abline(intercept = mm$coef[1], slope = mm$coef[2], color = "black",
              size = 1.2) +
  geom_abline(intercept = mf$coef[1], slope = mf$coef[2], color = "black",
```

```
            size = 1.2, alpha = 0.8) +
  labs(x = "Body mass (g)", y = "Brain mass (g)") +
  theme_bw() +
  ggtitle("Allometry of bird brains")
```



Analysis methods and result

Based on these analyses and results, we could write the Analysis Methods and Results as follows.

Analysis Methods We expected brain size to scale with body size according to a power-law relationship on the form $brainmass = a \times bodymass^b$. We linearized the expected power relationship through the logarithmic transformation $log(brainmass) = log(a) + b \times log(bodymass)$, and then fitted a linear regression model to the data. To assess whether the allometric slope (b) differs between the sexes, we analysed data for males and females separately.

Results Brain size scaled allometrically with body size (Fig. 1). In males, brain size increased by 5.5% per 10% increase in body mass (allometric slope = 0.55 ± 0.010), and body mass explained 83.5% of the variance in brain mass. The allometric slope was slightly steeper in females (0.57 ± 0.012).