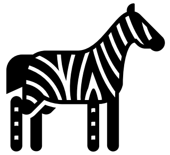
Tutorial 1
Basic statistical modelling in R

Clarke van Steenderen
Department of Zoology and Entomology
Rhodes University
South Africa

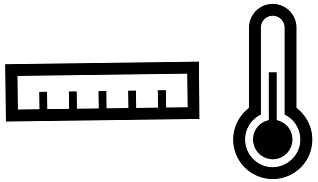
Clarke.vansteenderen@ru.ac.za

Data types vary greatly:

- **Counts → how many species/individuals**



- **Binary → one of two states → male/female, dead/alive**
- **Continuous → measurements such as height or mass**

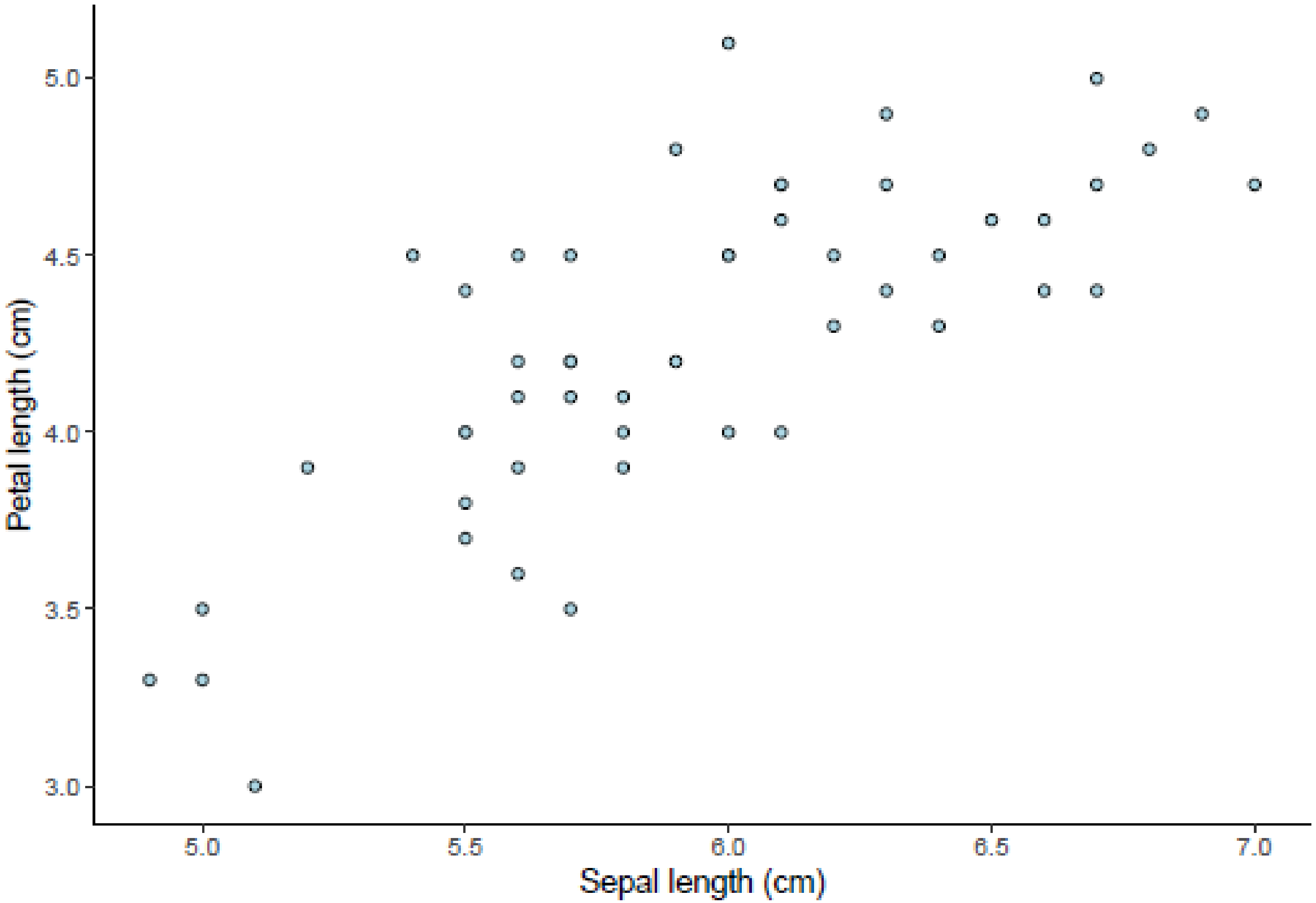


- **Proportions → proportion of a population with a disease**
- **Categorical → flowers classed into species or colours**

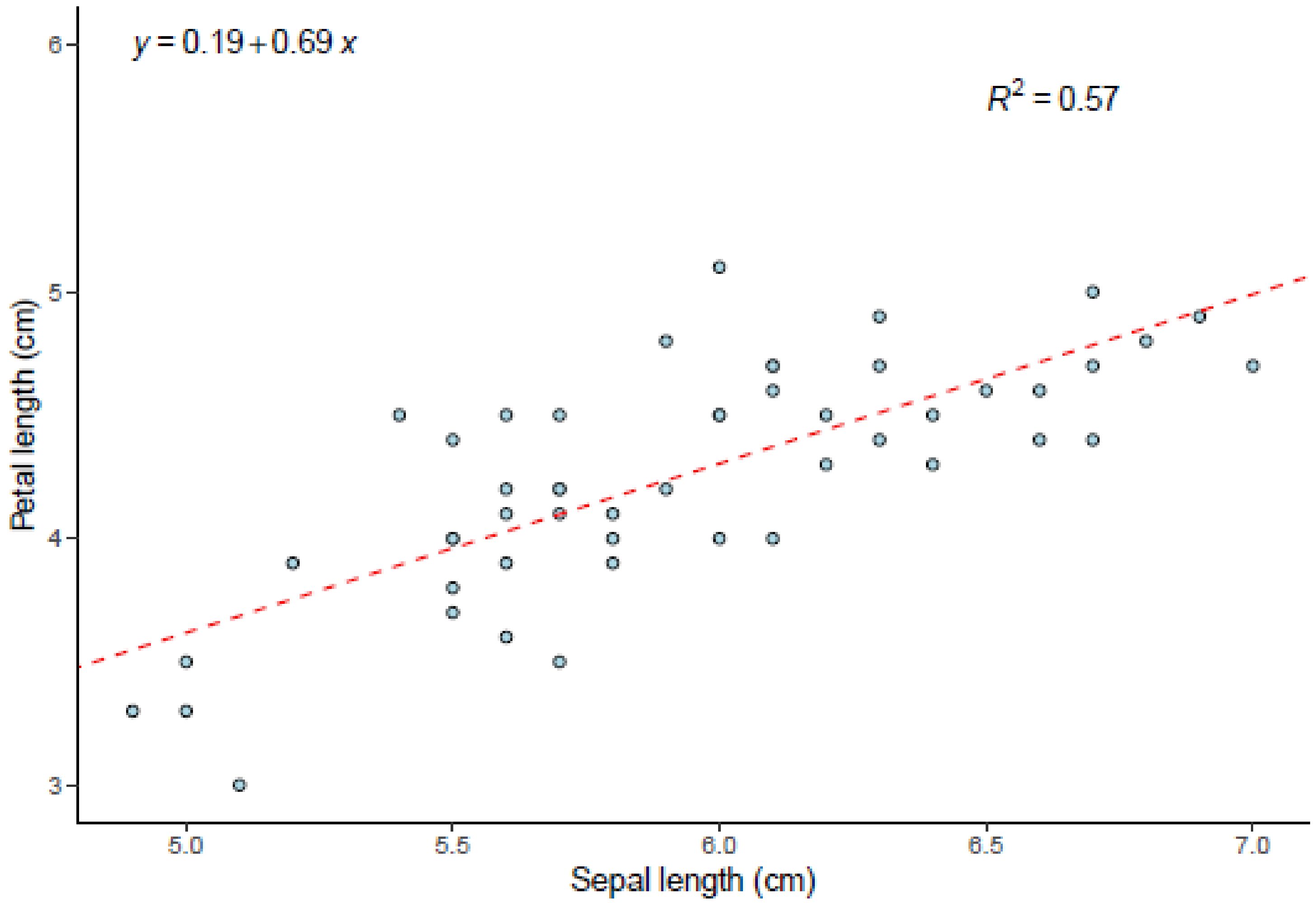
The type of data you have will determine the best choice of statistical test

Linear (straight line!) regression is useful to find whether there is:

- **A relationship between X (a predictor variable) and Y (a response variable) → e.g. does body weight (X) affect longevity (Y)? Is there a + or – relationship?**
- **Does a change in X lead to a significant difference in Y?**
- **Can X be used to accurately predict new values of Y?**

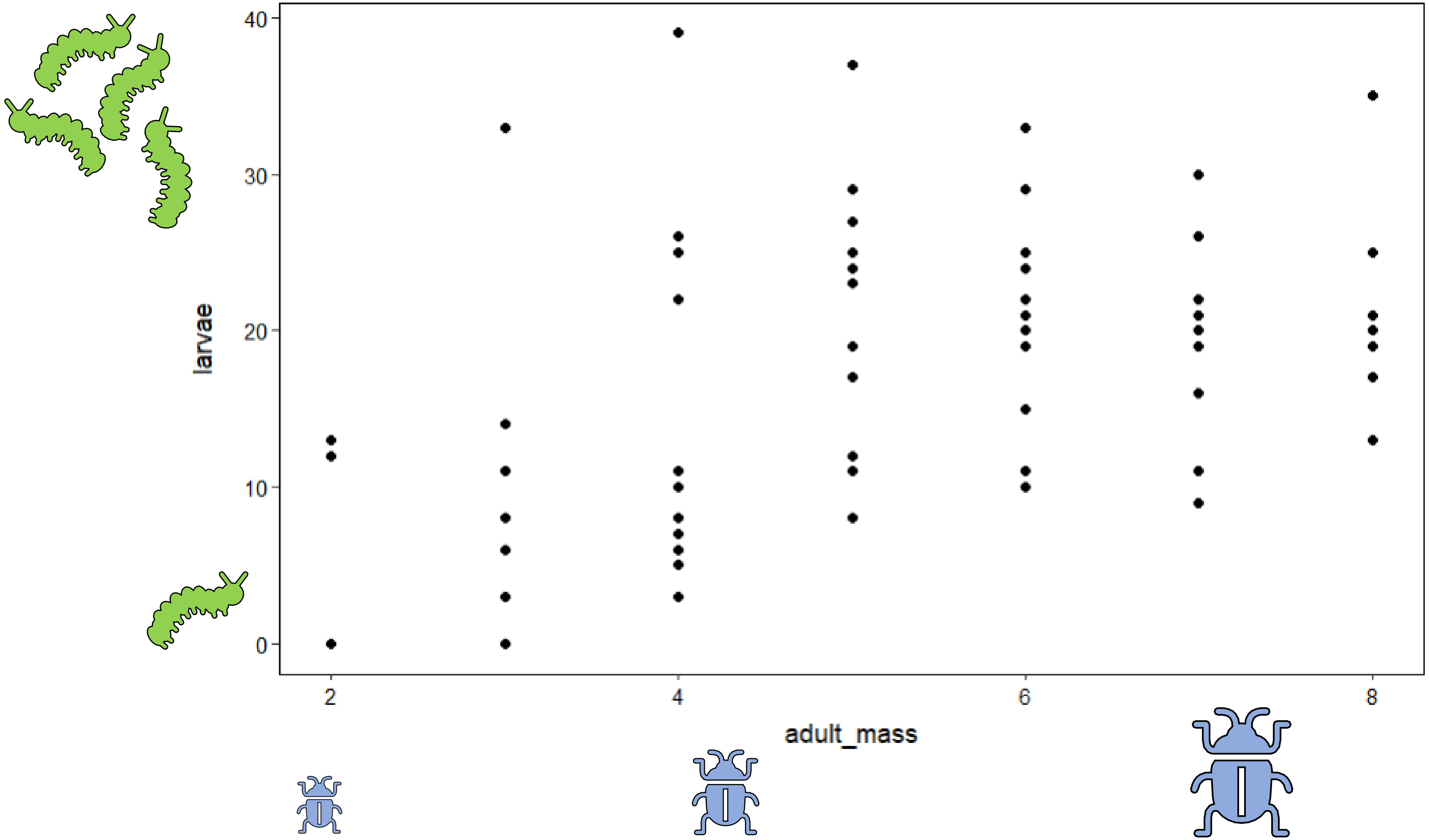


Petal length (cm)
5.0
4.5
4.0
3.5
3.0
5.0
5.5
6.0
6.5
7.0
Sepal length (cm)

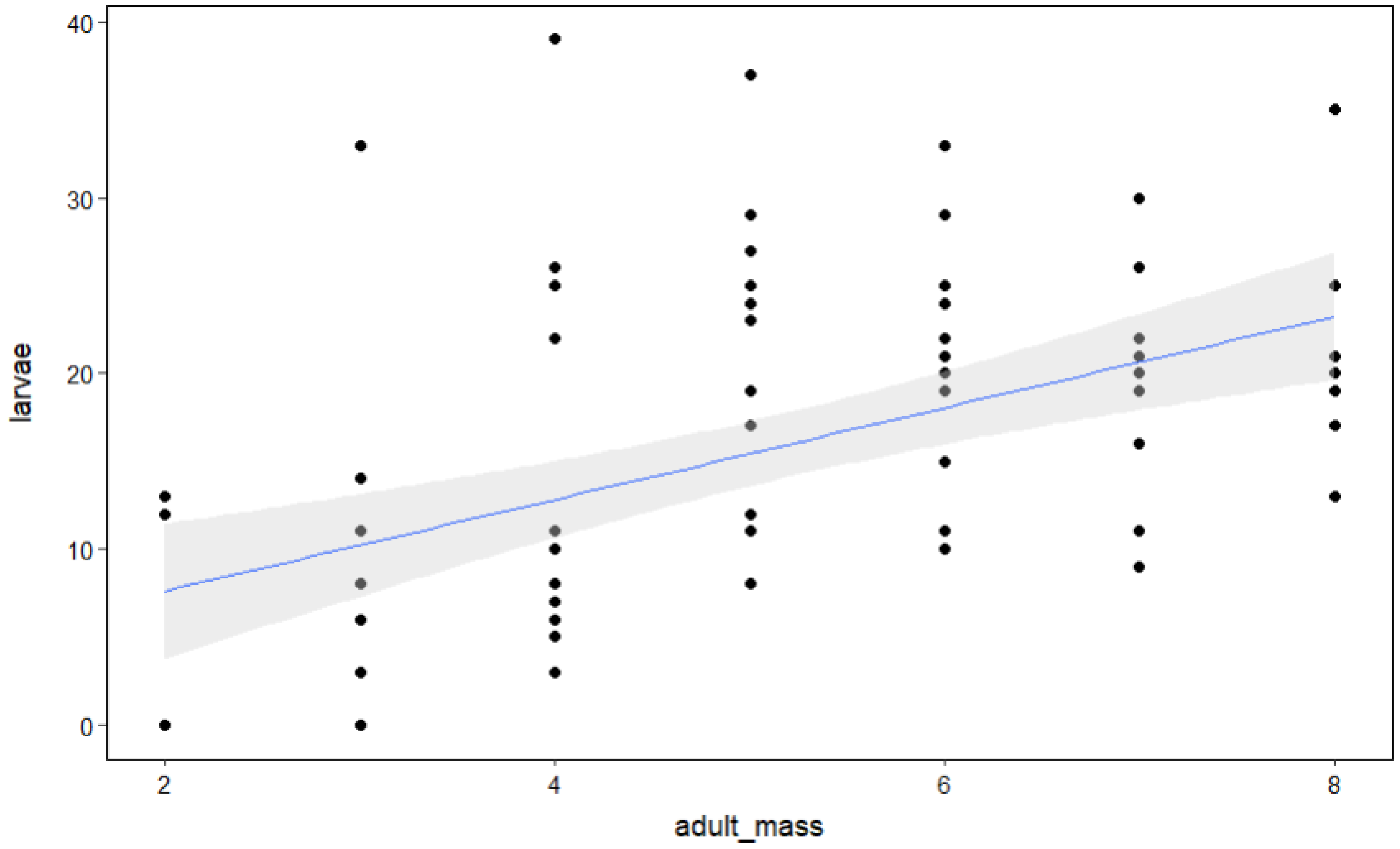


y = 0.19 + 0.69 x
R2 = 0.57
Petal length (cm)
Sepal length (cm)
6
5
4
3
5.0
5.5
6.0
6.5
7.0

What is the relationship between adult mass in an insect, and its reproductive output (number of larvae)? Positive, negative, none?



If we fit a straight line through these points, how well does it capture the trend? What is the variation in the data?



A linear regression line takes the form:

$$y = \beta_0 + \beta_1(x) + \varepsilon$$

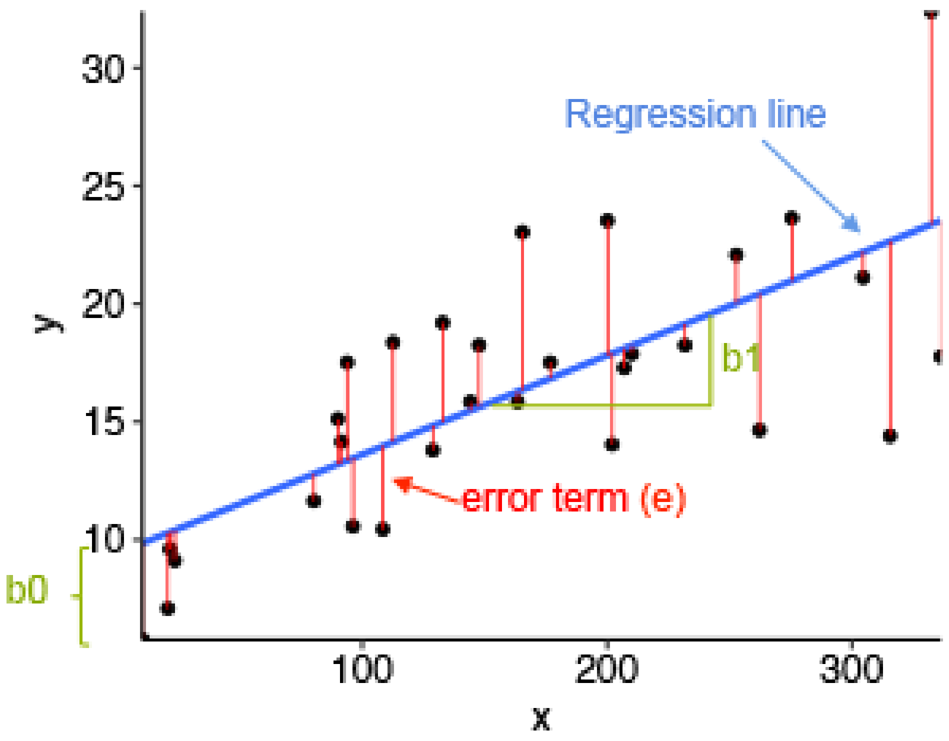
This should remind you of the equation of a straight line:

$$y = c + mx$$

- β_0 is the y-intercept
- β_1 is the gradient of the line
- ε is the error term → what is the difference between the actual measured y-values and the fitted line? This gives an indication of how much of the variation in the y-values is not captured by the model
- The β values are termed "Beta coefficients"

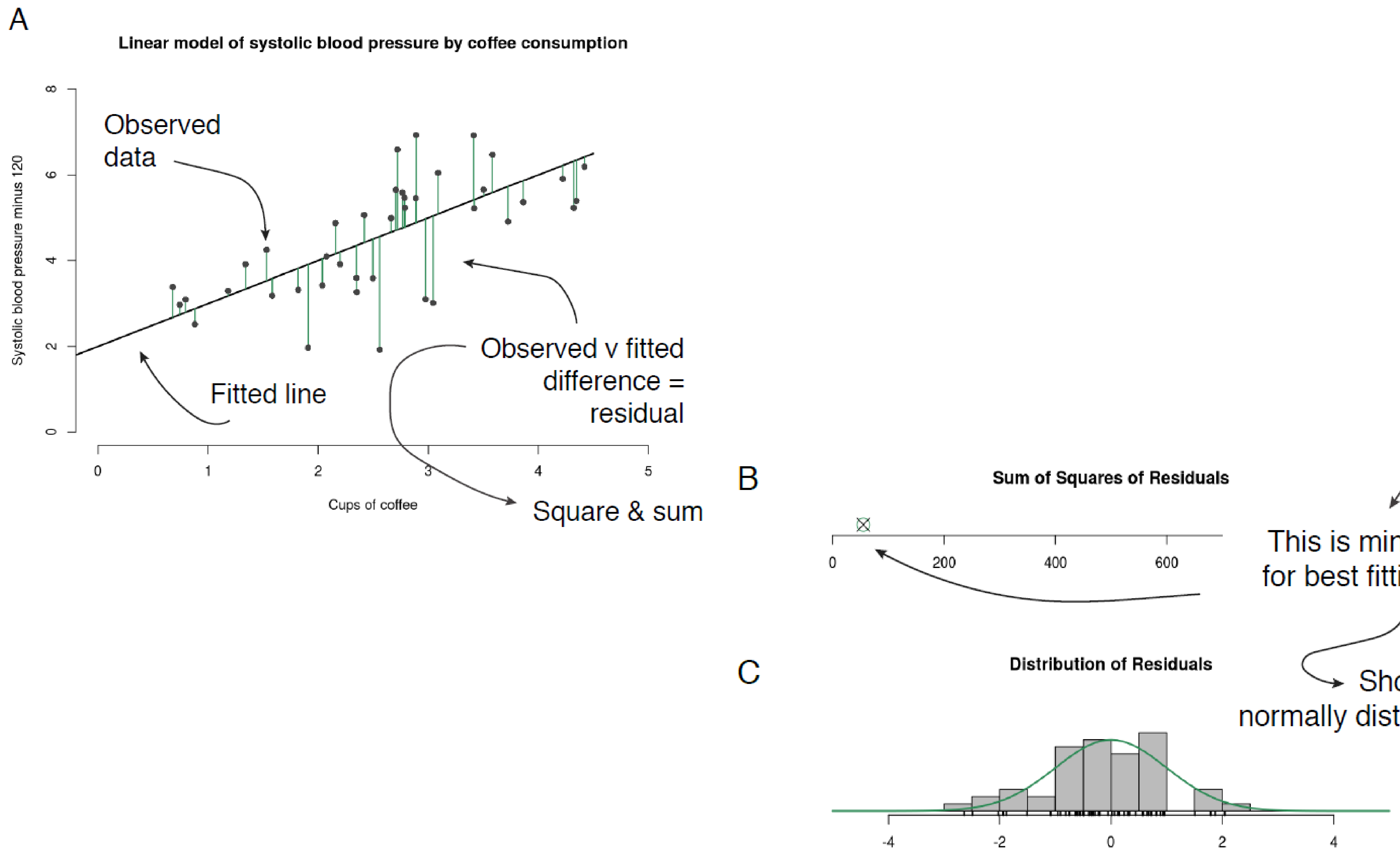
$y = \beta_0 + \beta_1(x) + \varepsilon$

$larvae = \beta_0 + \beta_1(adult\_mass) + \varepsilon$



The average variation of points around the regression line is the Residual Standard Error (RSE). The lower the RSE, the better the fit (ideal value is close to 0). Measured in the units of the dependent variable (e.g. # of larvae)

The R^2 value (between 0 and 1) indicates how well the model fits the data. 1 = excellent, 0 = no fit at all (measured as a proportion)



A
Linear model of systolic blood pressure by coffee consumption
Systolic blood pressure minus 120
Cups of coffee
Observed data
Fitted line
Observed v fitted difference = residual
Square & sum
B
Sum of Squares of Residuals
This is minimised for best fitting line
C
Distribution of Residuals
Should be normally distributed

Model fitting in R

```
linear.mod.1 = lm(y ~ x, data = data)
linear.mod.1 = lm(larvae ~ adult_mass, data = in.data)
```

- `linear.mod.1`: **the name in which we are storing the model → this can be any name you choose, but make it informative**
- `lm`: **the linear model function in R**
- `Y~X` **means Y as a function of X → how does Y change with a change in X? Note the use of the tilde sign (~)**
- `in.data` **is the input data that is read into R (e.g. from an Excel sheet)**

Check how well the model fits

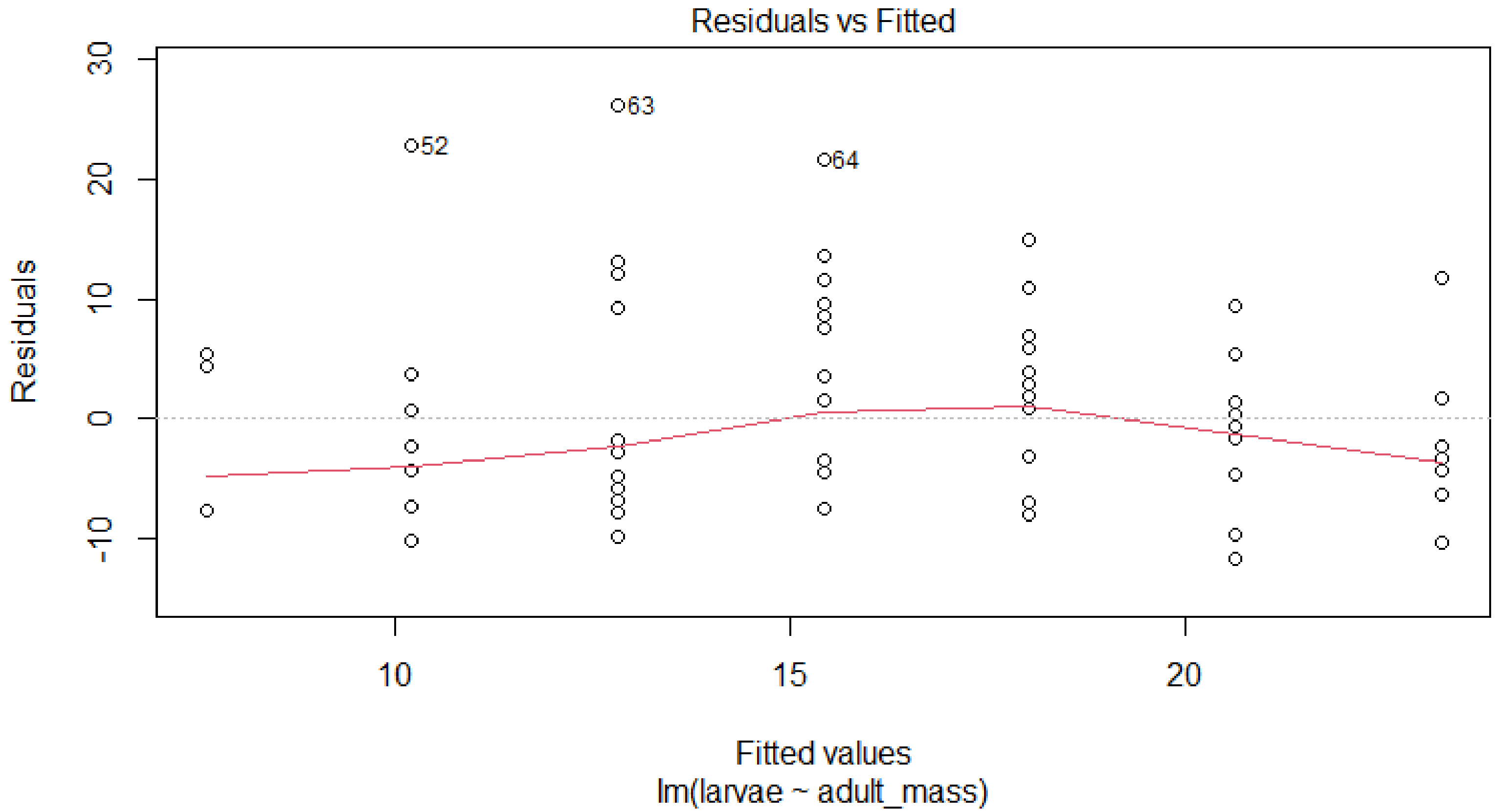
For a statistical model to be reliable, there are a few assumptions that the data need to meet:

1. **Linearity**
2. **Independence → data collected from one site/individual does not affect data collected in others**
3. **Normality of residuals**
4. **Equal variance**

We need to run model diagnostic tests to check these. In R, the `DHARMa` package is very useful

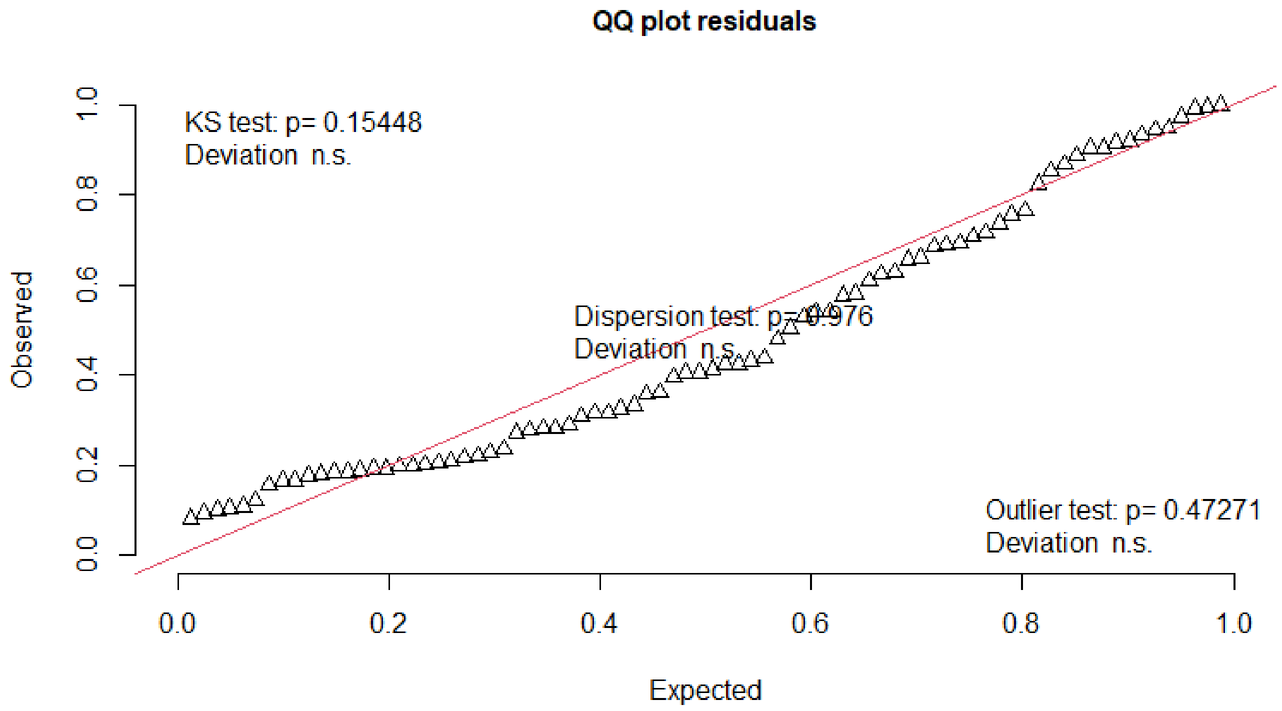
1. Linearity → residual vs fitted plot

Residual values should cluster around the y = 0 line, with no clustering or patterning



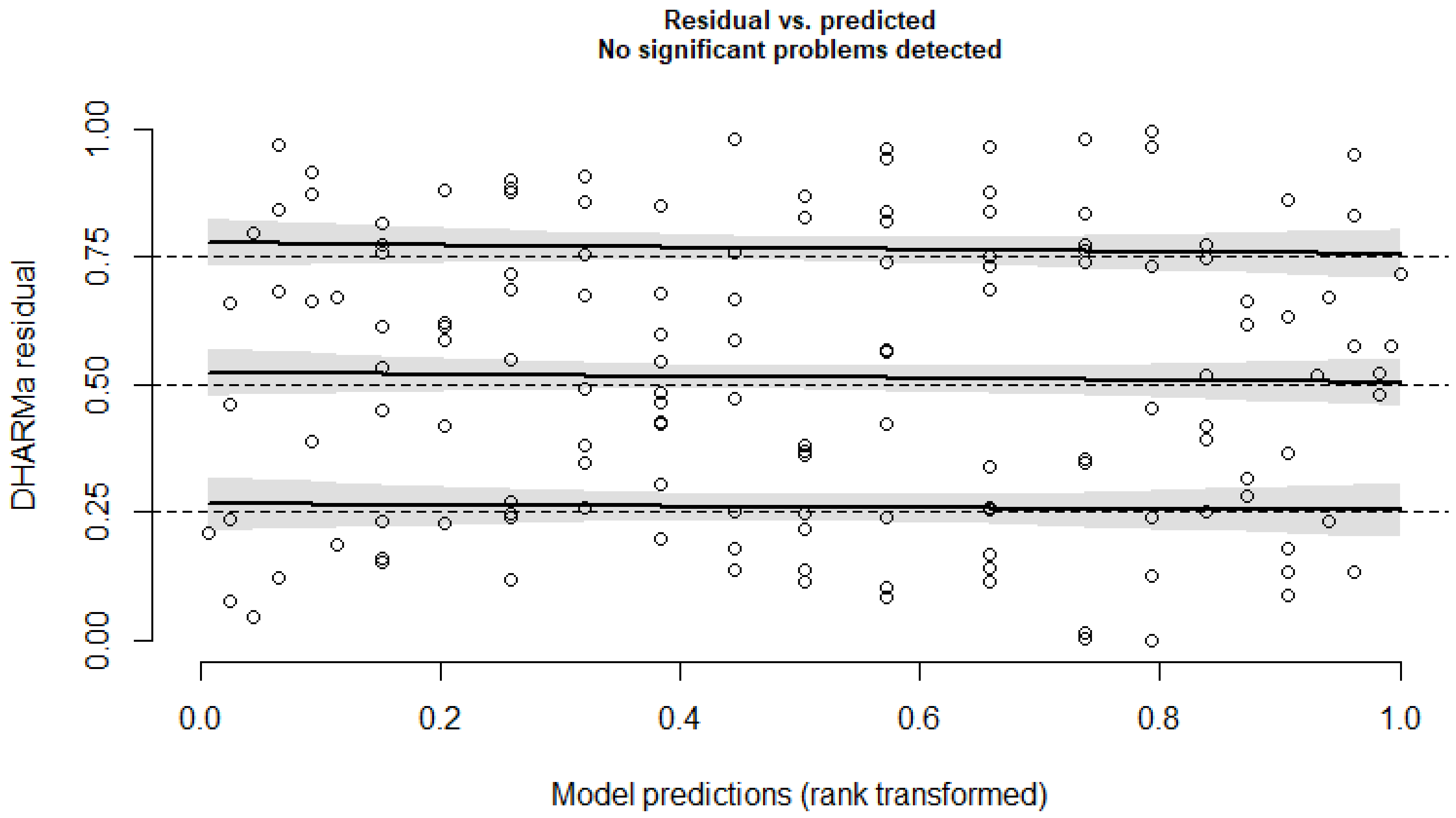
3. Normality of residuals → QQ plot

Residual values should cluster around the straight line with gradient of 1 (red line). Kolmogorov-Smirnov test shows whether residuals are normally distributed ($p > 0.05$) or not ($p < 0.05$). Dispersion and outlier tests should ideally be non-significant



4. Equal variance → residual vs predicted

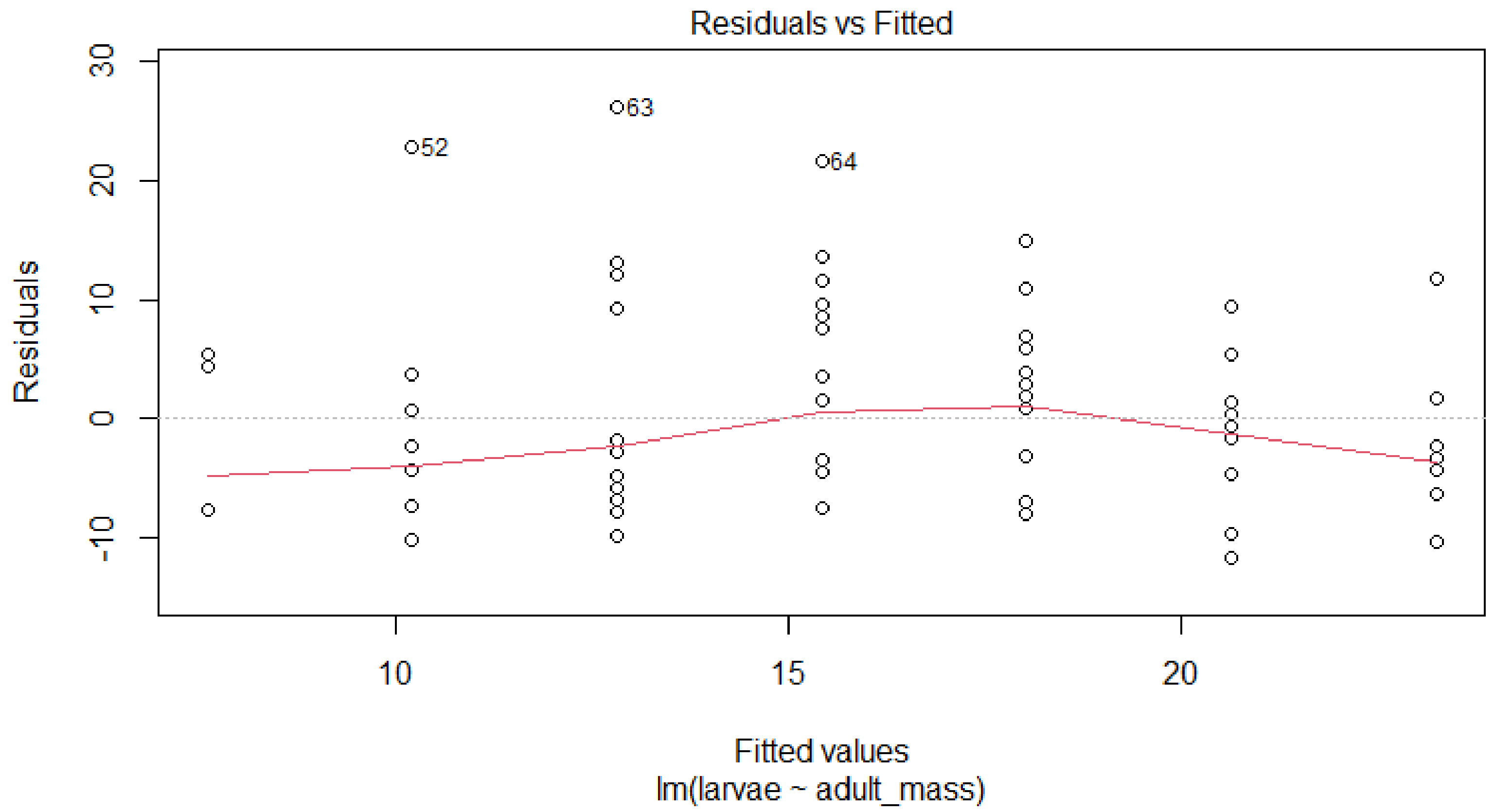
Bold lines should fall along the y = 0.25, 0.5, and 0.75 quartile range marks



Check how well the model fits

```
plot(linear.mod.1, which = 1)
```

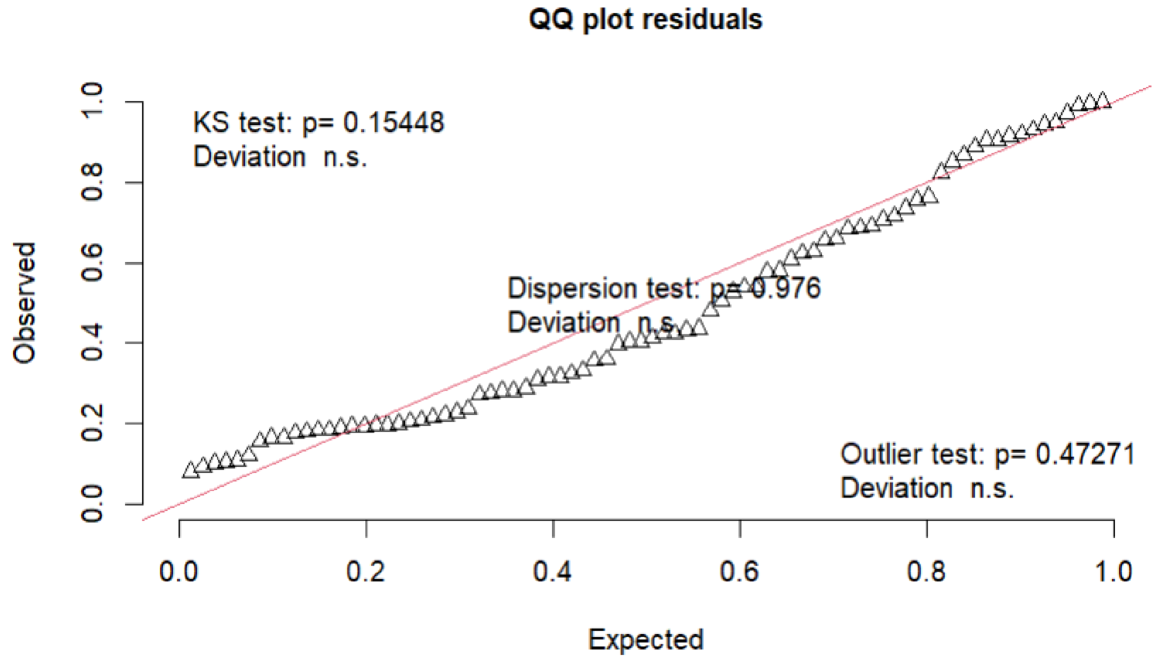
Here we see an approximate straight red line along the y = 0 line



Check how well the model fits

`DHARMa::plotQQunif(linear.mod.1)`

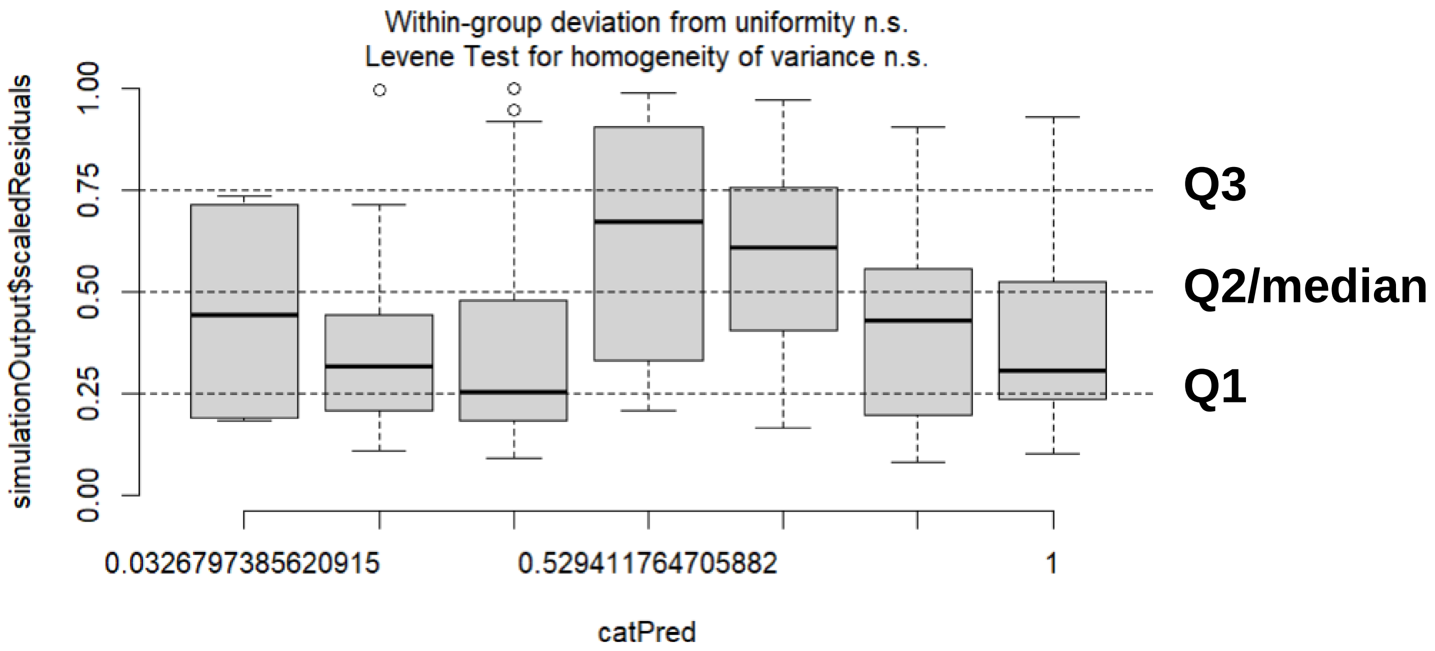
This QQ plot shows that residuals are normally distributed (KS test), that there is no overdispersion, and no outliers. We want to see the points roughly matching the straight red line (y = x)



Check how well the model fits

```
DHARMa::plotResiduals(linear.mod.1)
```

Within-group deviation (Kolmogorov-Smirnov test (KS)) per mass category and Levene tests for homogeneity of variance show that there are no issues with the residuals in this linear model. We expect the three interquartile ranges of the boxplots to match the dotted horizontal lines



Check how well the model fits

```
summary(linear.mod.1)
```

The summary output for our model shows that our intercept (β_0) = 2.399, and our gradient (β_1) is 2.6066. Essentially, our straight line is y = 2.6x + 2.4

```
Call:
lm(formula = larvae ~ adult_mass, data = in.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.645  -6.825  -2.022   4.068  26.175 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) β0 2.3990     2.9254   0.820    0.415    
adult_mass  β1 2.6066     0.5431   4.799 7.52e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.236 on 78 degrees of freedom
Multiple R-squared:  0.228,	Adjusted R-squared:  0.2181 
F-statistic: 23.03 on 1 and 78 DF,  p-value: 7.521e-06
```

```
Call:
lm(formula = larvae ~ adult_mass, data = in.data)

Residuals:
    Min      1Q  Median      3Q     Max
-11.645  -6.825  -2.022   4.068  26.175

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) β0 2.3990     2.9254   0.820    0.415
adult_mass  β1 2.6066     0.5431   4.799 7.52e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.236 on 78 degrees of freedom
Multiple R-squared:  0.228,	Adjusted R-squared:  0.2181
F-statistic: 23.03 on 1 and 78 DF,  p-value: 7.521e-06
```

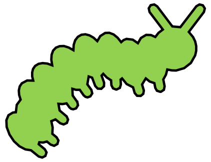
- **β_1 suggests that for every unit increase in biomass (1 g), the number of larvae increase by 2.6**
- **Our p-value is < 0.001, which means that adult mass has a significant effect on larvae number**
- **RSE tells us that adult mass (x) predicts larvae numbers (y) with an average error of 8.2 larvae**
- **The R^2 value indicates that our linear model explains 0.22 (22%) of the variation in the data.**
  What might this indicate?

Check how well the model fits

```
confit(linear.mod.1)
```

```
                2.5 %    97.5 %
(Intercept)   -3.424936 8.222914
adult_mass     1.525312 3.687797
```

The 95% confidence interval suggests that a 1 g increase in adult mass will result in 1.5 – 3.7 more individual larvae. If we were to repeat this experiment 100 times, our larvae estimates will fall in this range 95 times out of the 100.



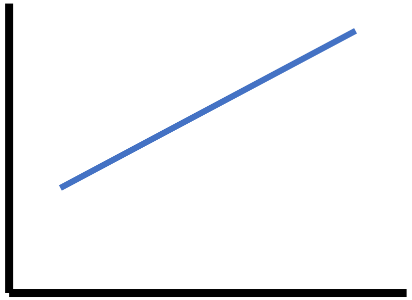
Hypothesis testing

- **Null (H_0)**

Body mass does not have a significant effect on reproductive output (number of larvae)

- **Alternative (H_1)**

Body mass significantly influences reproductive output



Hypothesis testing

Instead of lm(), we will run a glm() → more on that later

```
H0.model = glm(larvae ~ 1, data = in.data,
family = gaussian)

H1.model = glm(larvae ~ 1 + adult_mass,
data = in.data, family = gaussian)
```

A Gaussian distribution is specified for normally-distributed residuals

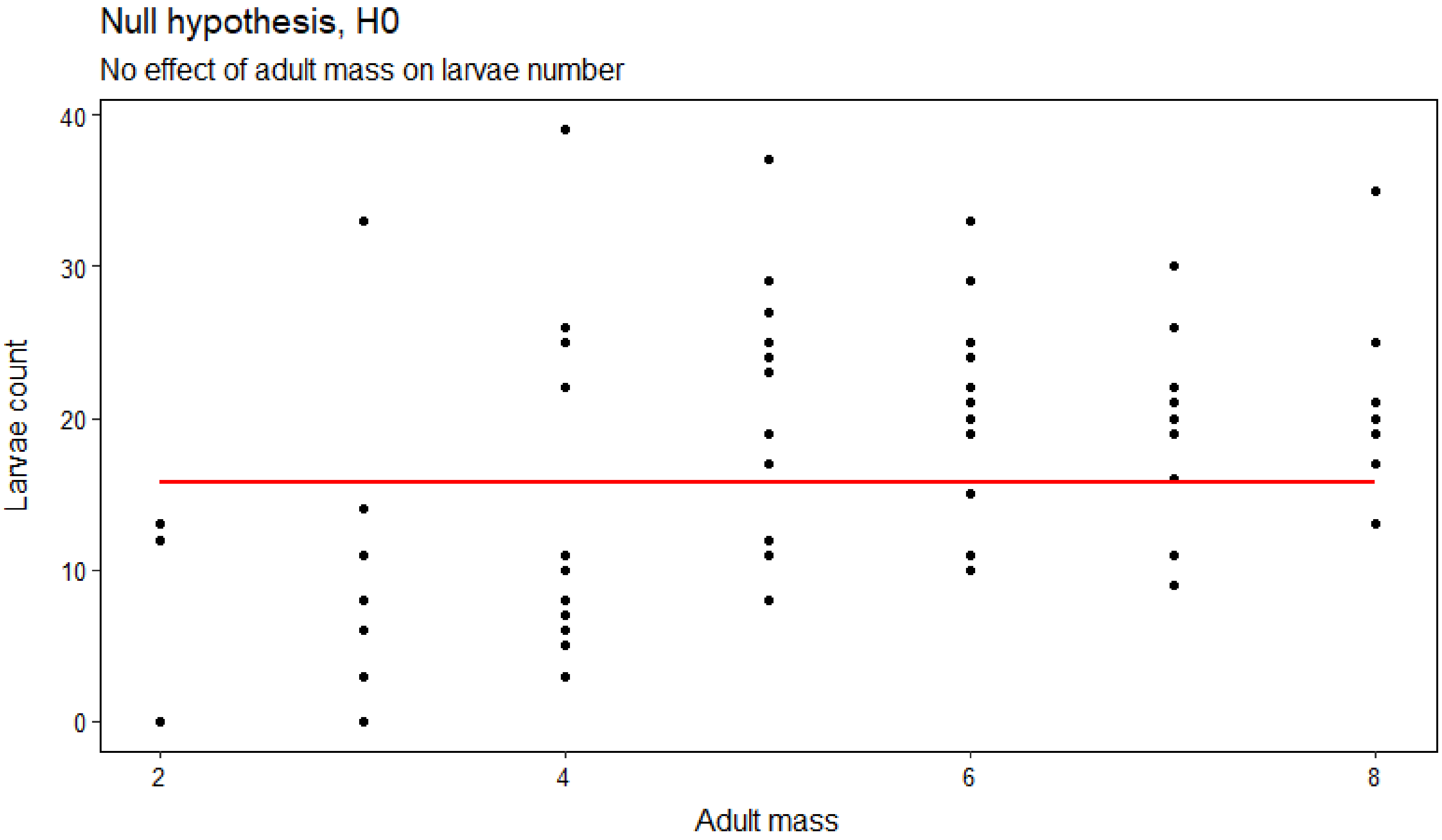
Notice how the `H0.model` excludes adult mass, and only runs the model with an intercept term (~1). We are not including any predictor variables here, and are assuming that β_1 (gradient) is not significantly different from zero (i.e. a horizontal line)

```
H0.model = glm(larvae ~ 1, data = in.data,
family = gaussian)

H1.model = glm(larvae ~ 1 + adult_mass,
data = in.data, family = gaussian)
```

summary(H0.model)

```
> summary(H0.model)

Call:
glm(formula = larvae ~ 1, family = gaussian, data = in.data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.725      1.041    15.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 86.75886)

    Null deviance: 6853.9  on 79  degrees of freedom
Residual deviance: 6853.9  on 79  degrees of freedom
AIC: 587.07

Number of Fisher Scoring iterations: 2
```

$y = 15.73$

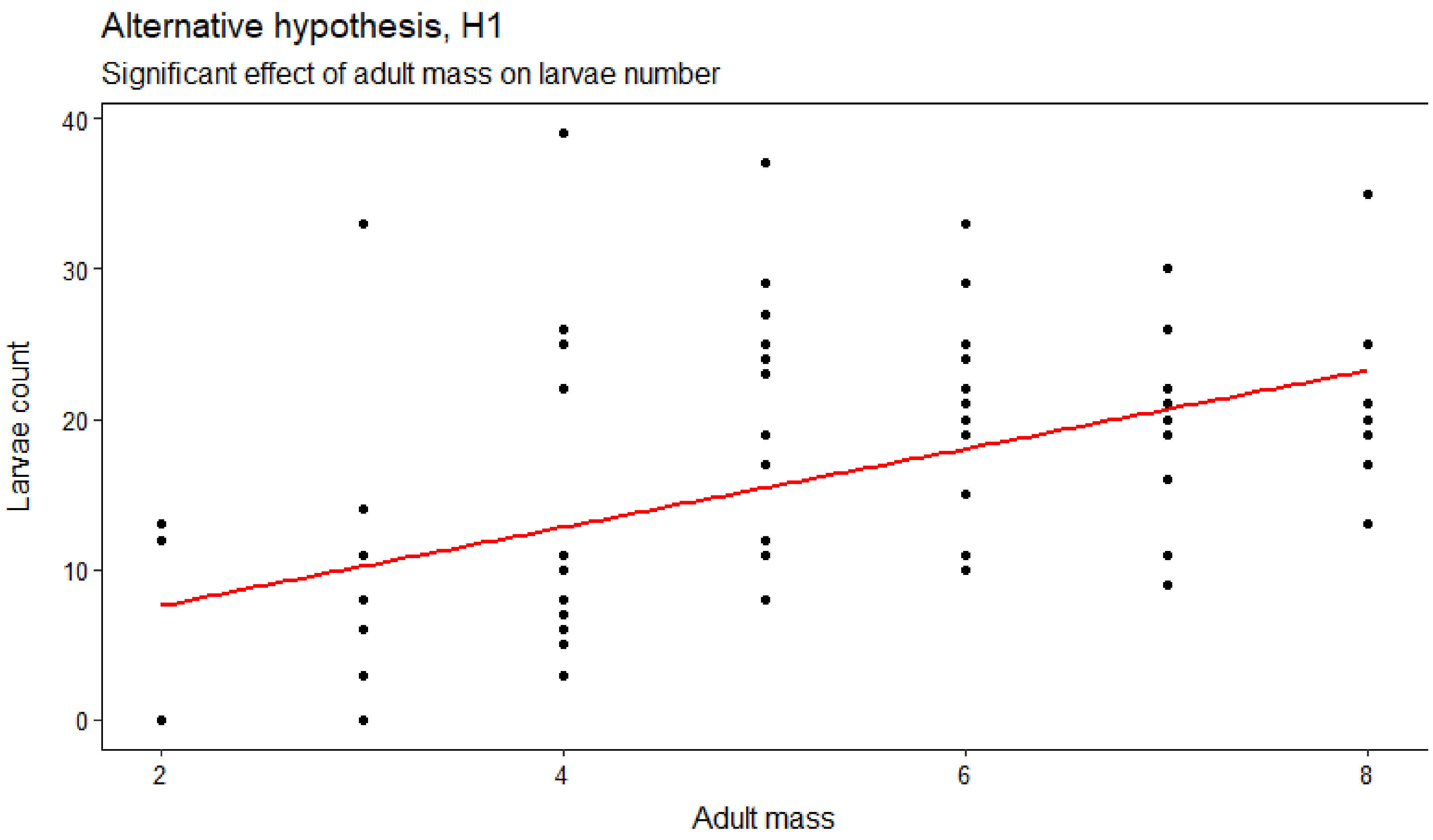


Null hypothesis, H0
No effect of adult mass on larvae number
Larvae count
40
30
20
10
0
2
4
6
8
Adult mass

summary(H1.model)

```
Call:
glm(formula = larvae ~ 1 + adult_mass, family = gaussian, data = in.data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3990     2.9254   0.820    0.415
adult_mass    2.6066     0.5431   4.799 7.52e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 67.83828)

    Null deviance: 6853.9  on 79  degrees of freedom
Residual deviance: 5291.4  on 78  degrees of freedom
AIC: 568.37

Number of Fisher Scoring iterations: 2
```

$$y = 2.6x + 2.4$$



Alternative hypothesis, H1
Significant effect of adult mass on larvae number
Larvae count
0
10
20
30
40
2
4
6
8
Adult mass