# Session #1: Linear Models

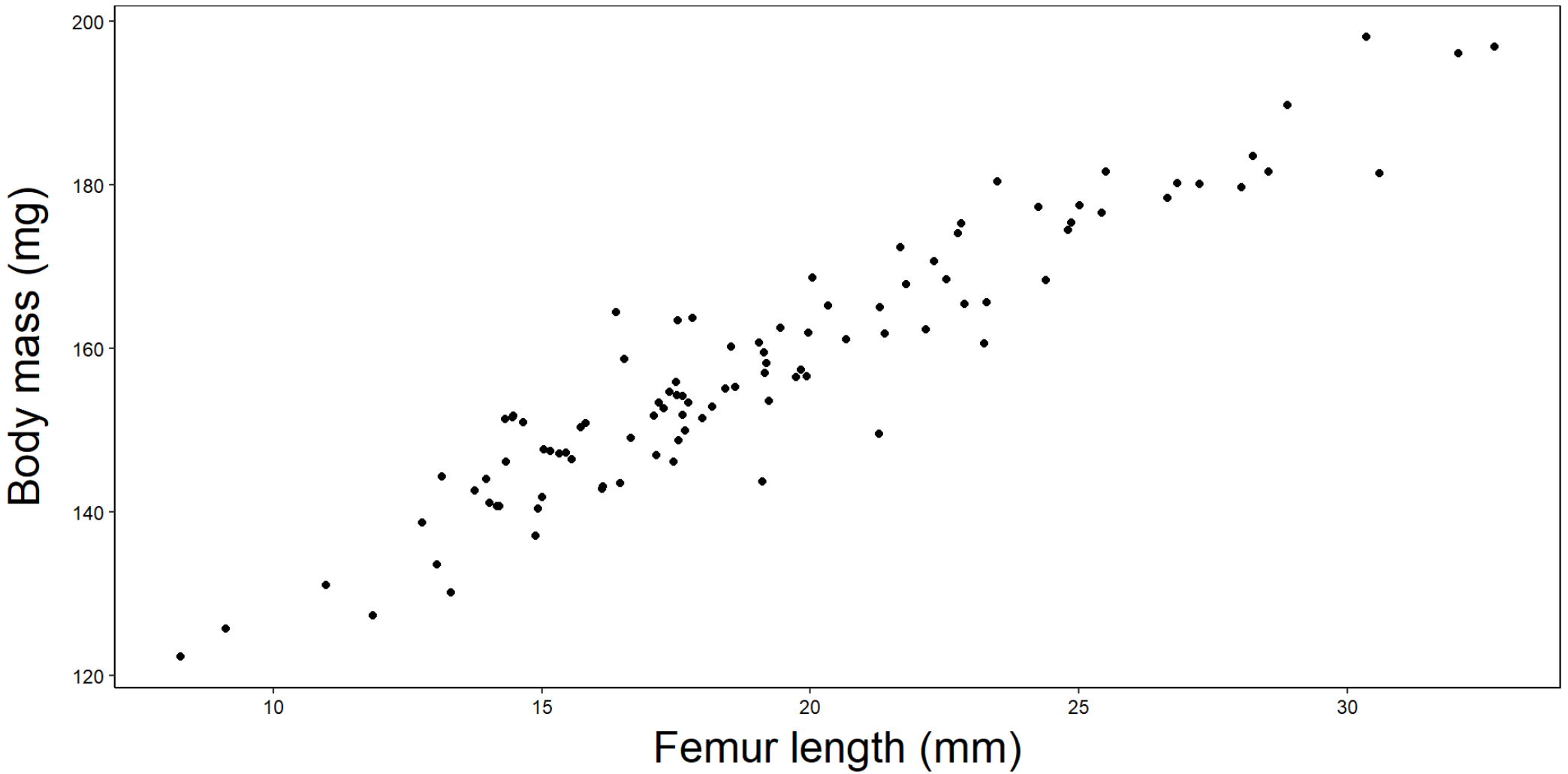
Guy F. Sutton
Centre for Biological Control
Rhodes University, South Africa
Email: g.sutton@ru.ac.za

# Linear Models

- Aims of linear regression
  - 1. Is there a relationship/correlation between Y and X?
  - 2. Is there a difference in Y due to the values of X?
  - 3. Can I use X to predict Y?

# Linear regression basics

Is female femur length correlated with body size?



# Response variable

- What is our response/dependent variable?
  - The measure we are interested in predicting or explaining
    - Body mass (`bodyMass`)
    - Usually denoted by `y` or `Y`

# Predictor variable

- What is our covariate/fixed effect/predictor variable?
  - The variable used to explain the variation/differences observed in the response variable
    - Femur length (`femurLength`)
    - Usually denoted by `x`, `X`, or `Xn`

# Model fitting

# Linear model formula

In R, we can fit a simple linear model using:

```
m1 <- glm(y ~ x, data = data)
```

- Where:
  - `m1` is a object name where we store our model
  - `glm` is a built-in function to run a linear model
  - `data` is the name of the dataset containing our data
  - `y` is the column name in `data` containing our response variable
  - `x` is the column name in `data` containing our predictor variable

# Insert data object name

## The data is stored in an object called df

```
head(df, 10)
```

```
# A tibble: 10 × 2
   femurLength bodyMass
         <dbl>    <dbl>
 1        14.0      144.
 2        21.4      162.
 3        25.4      177.
 4         8.27     122.
 5        22.1      162.
 6        22.5      168.
 7        17.1      147.
 8        17.3      153.
 9        17.2      153.
10        15.5      146.
```

```
m1 <- glm(
  y ~ x,
  data = df
)
```

# Insert response variable name

## The response variable is stored in a column called `bodyMass`

```
head(df, 10)
```

```
# A tibble: 10 × 2
   femurLength bodyMass
         <dbl>    <dbl>
 1        14.0      144.
 2        21.4      162.
 3        25.4      177.
 4         8.27     122.
 5        22.1      162.
 6        22.5      168.
 7        17.1      147.
 8        17.3      153.
 9        17.2      153.
10        15.5      146.
```

```
m1 <- glm(
  bodyMass ~ x,
  data = df
)
```

# Insert predictor variable name

The predictor variable is stored in a column called `femurLength`

```
head(df, 10)
```

```
# A tibble: 10 × 2
   femurLength bodyMass
         <dbl>    <dbl>
 1        14.0      144.
 2        21.4      162.
 3        25.4      177.
 4         8.27     122.
 5        22.1      162.
 6        22.5      168.
 7        17.1      147.
 8        17.3      153.
 9        17.2      153.
10        15.5      146.
```

```
m1 <- glm(
  bodyMass ~ femurLength,
  data = df
)
```

# Research question

Simple linear regression modelling body mass as a linear function of femur length

- *Research Q*: Is there a correlation between `femurLength` and `bodyMass`?
  - I.e. What is the relationship between `femurLength` and `bodyMass`?
  - I.e. Do larger individuals weigh more?

```
m1 <- glm(
  bodyMass ~ femurLength,
  data = df
)
```

# Model equation

Simple linear regression modelling body mass as a linear function of femur length

```
m1 <- glm(
  bodyMass ~ femurLength,
  data = df
)
```

$$\text{bodyMass} = \beta_0 + \beta_1 (\text{femurLe}$$

# Global intercept

$$\text{bodyMass} = \beta_0 + \beta_1(\text{femurLe}$$

- $\beta_0$: Intercept
  - The expected value of Y (`bodyMass`) when X = 0 (`femurLength` = 0)

# Beta coefficient

$$\text{bodyMass} = \beta_0 + \beta_1(\text{femurLe}$$

- $\beta_1$ Beta coefficient / slope coefficient
  - The expected change in Y (`bodyMass`) for every unit-change in X (`femurLength`)
    - E.g. For every 1mm increase in `femurLength`, by how much does `bodyMass` change, on average?

# Error term

$$\text{bodyMass} = \beta_0 + \beta_1(\text{femurLe}$$

- $\epsilon$Error term
  - The difference between the actual Y-values (measured `bodyMass` values) and the expected value of Y based on the model we have fit
  - This effectively tells us how much of the observed variation in `bodyMass` is **NOT** due to the linear effect of `femurLength`

# Model summary - intercept

```r
m1 <- glm(bodyMass ~ femurLength, data = df)
summary(m1)
```

```
Call:
glm(formula = bodyMass ~ femurLength, data = df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.7075     2.0605   48.88   <2e-16 ***
femurLength   2.9739     0.1038   28.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 26.88972)

    Null deviance: 24717.8  on 99  degrees of freedom
Residual deviance:  2635.2  on 98  degrees of freedom
AIC: 616.94

Number of Fisher Scoring iterations: 2
```

- $\beta_0$ = 100.70
  - When `femurLength` = 0 mm, the expected `bodyMass` = 100.70mg
    - Does this make sense?

# Model summary - slope

```r
m1 <- glm(bodyMass ~ femurLength, data = df)
summary(m1)
```

```
Call:
glm(formula = bodyMass ~ femurLength, data = df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.7075     2.0605   48.88   <2e-16 ***
femurLength   2.9739     0.1038   28.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 26.88972)

    Null deviance: 24717.8  on 99  degrees of freedom
Residual deviance:  2635.2  on 98  degrees of freedom
AIC: 616.94

Number of Fisher Scoring iterations: 2
```

- $\beta_1$ = 2.97
  - The expected change in `bodyMass` for every 1-unit change in `femurLength`
    - E.g. For every 1mm increase in `femurLength`, `bodyMass` increases by 2.97mg, on average

# Model diagnostics

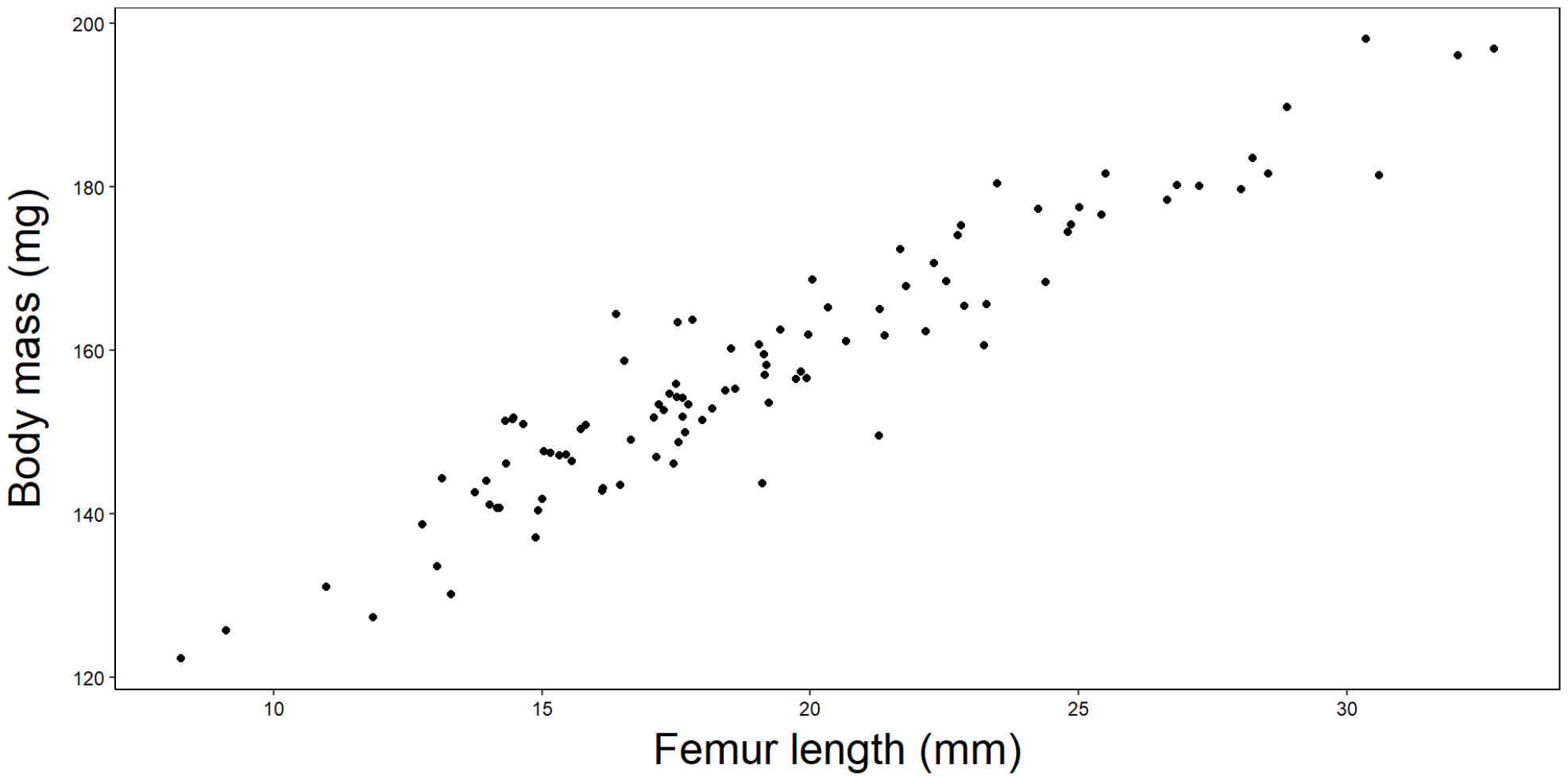
# Model diagnostics

For our model and any inferences to be valid, we need to check that it meets a few assumptions:

1. `Linearity`: Linear relationship between Y (`bodyMass`) and X (`femurLength`)
2. `Independence`: Data points are independent (no connection or unaccounted for clustering of data)
3. `Normality`: Responses are normally distributed for each level in X
4. `Equal variance`: The variation in responses are equal for each level in X

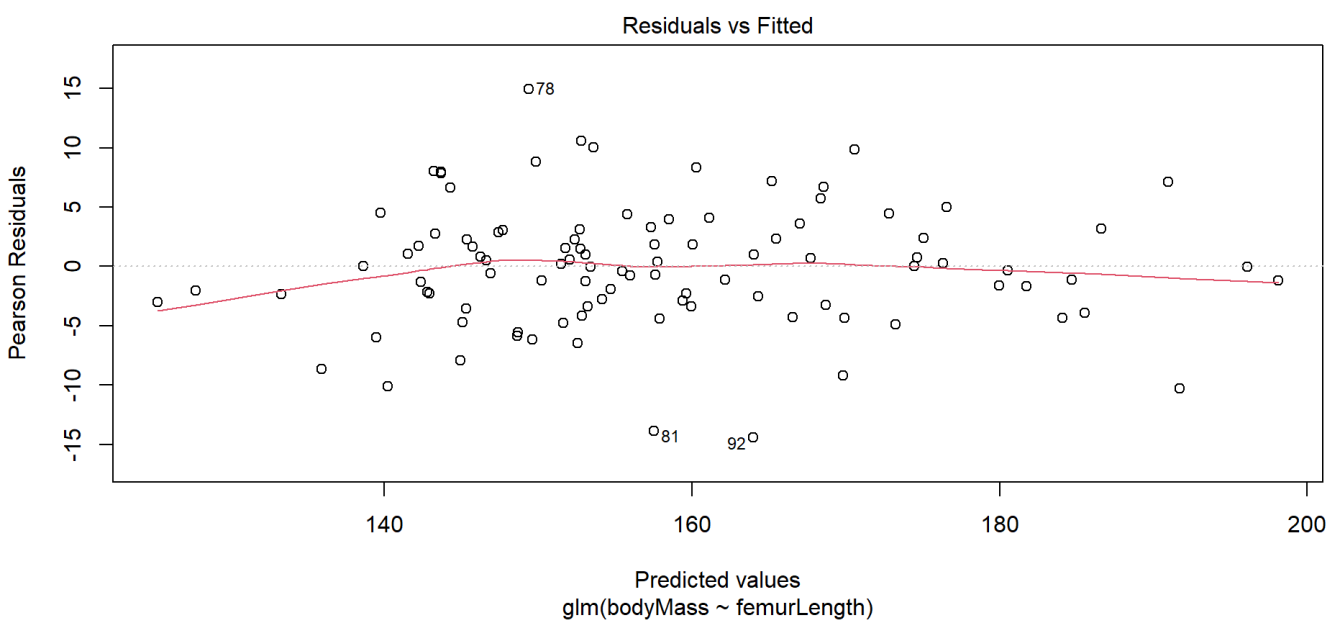
# 1. Linearity

Looks like there is a linear relationship between `bodyLength` and `femurLength`.



# 1. Linearity

- Need to confirm that our model was specified correctly to capture this relationship
  - Use a `residual vs fitted plot`
  - Residuals should cluster around y = 0, no pattern evident



# 2. Independence

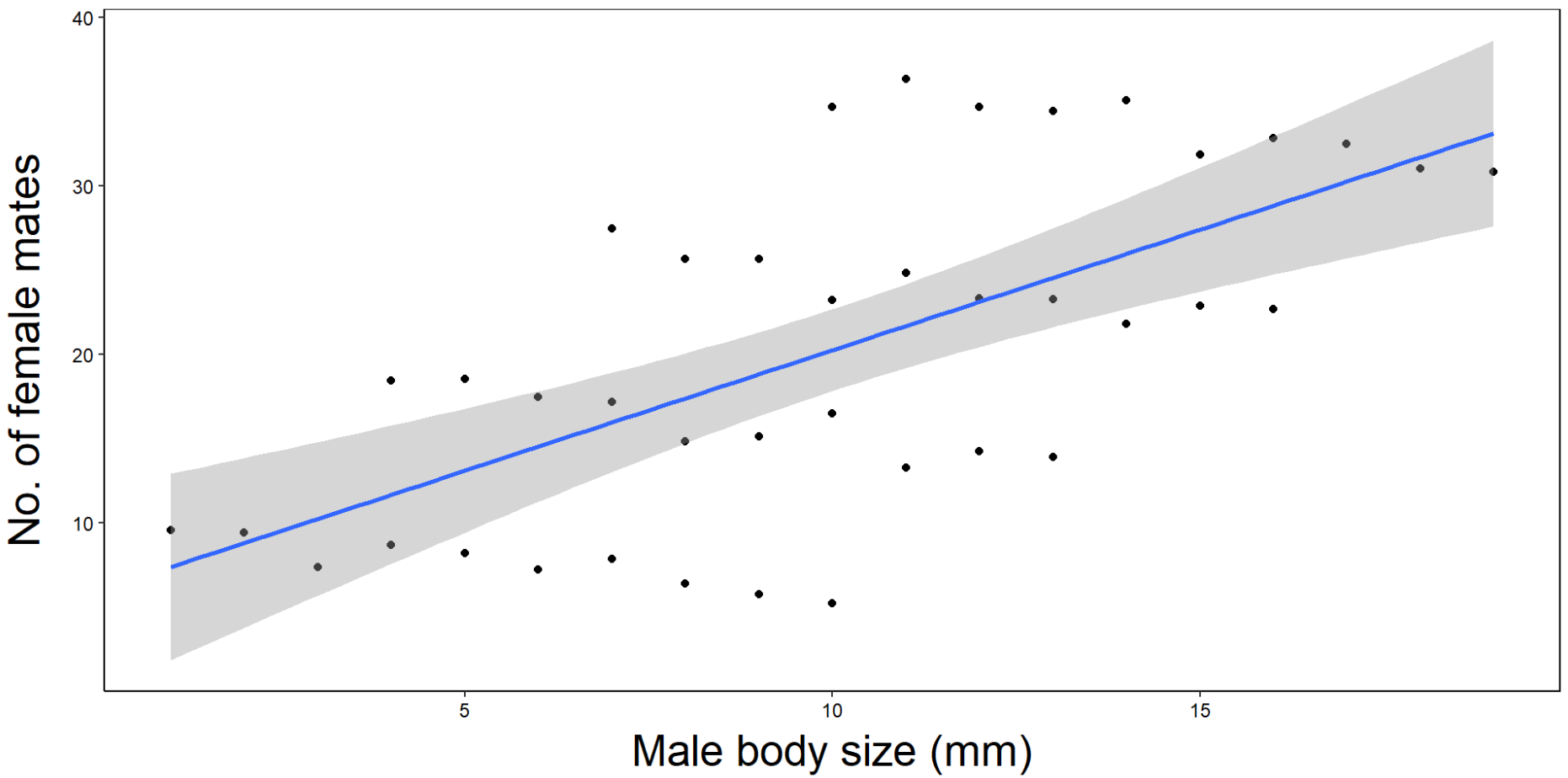
- Independence is almost impossible to diagnose with a plot or statistical test.
  - Independence is a property of the experimental design
  - Observations are not grouped or clustered, not more similar to each other in a systematic way

# 2. Independence

- Let’s consider a study where we observe 40 male wasps and count how many females they mate with over a 10 minute period?
  - Do larger males attract more mates?

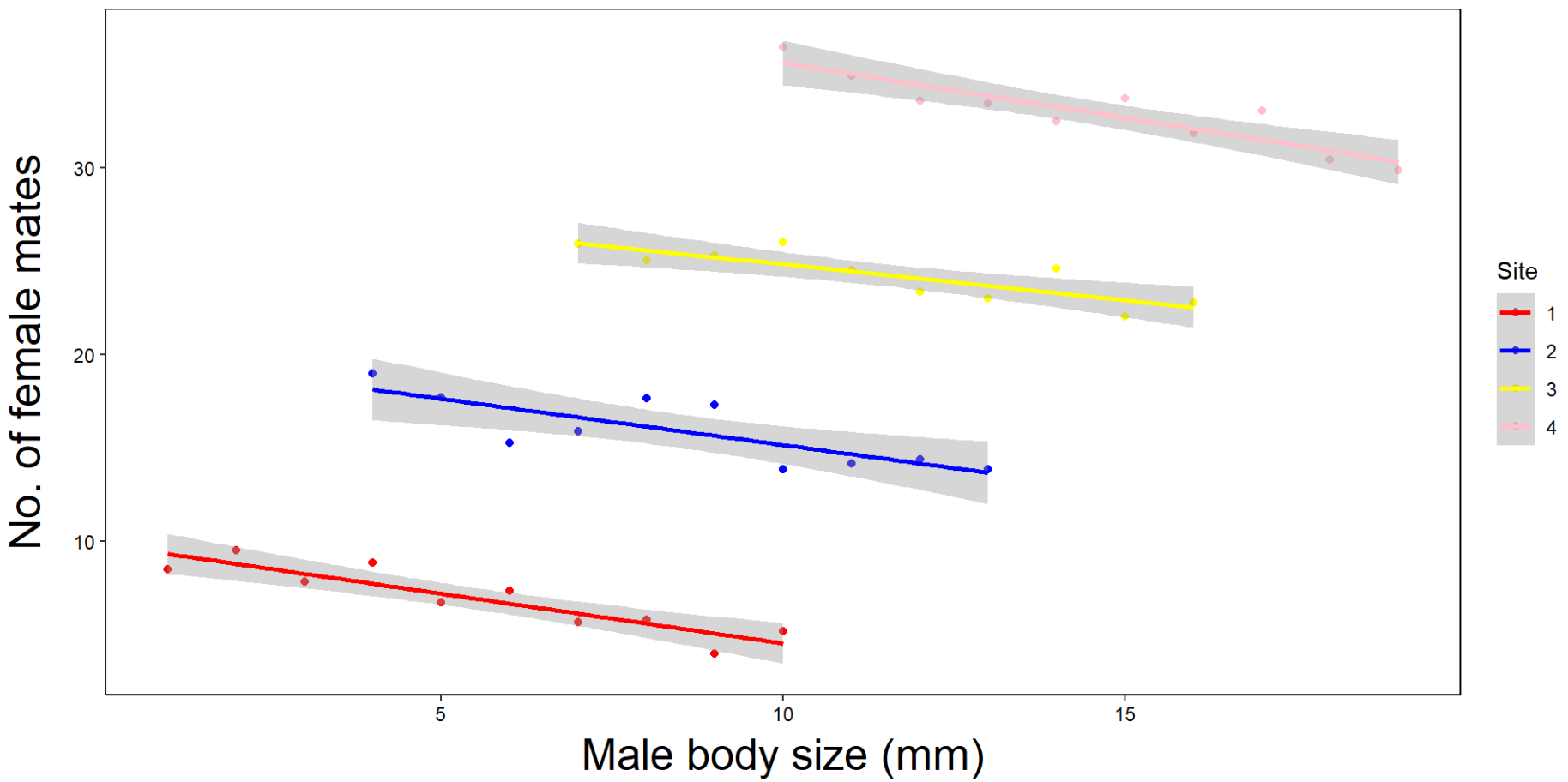
# 2. Independence

- If we correlate male body size with no. of mates, there is a clear positive relationship, right?



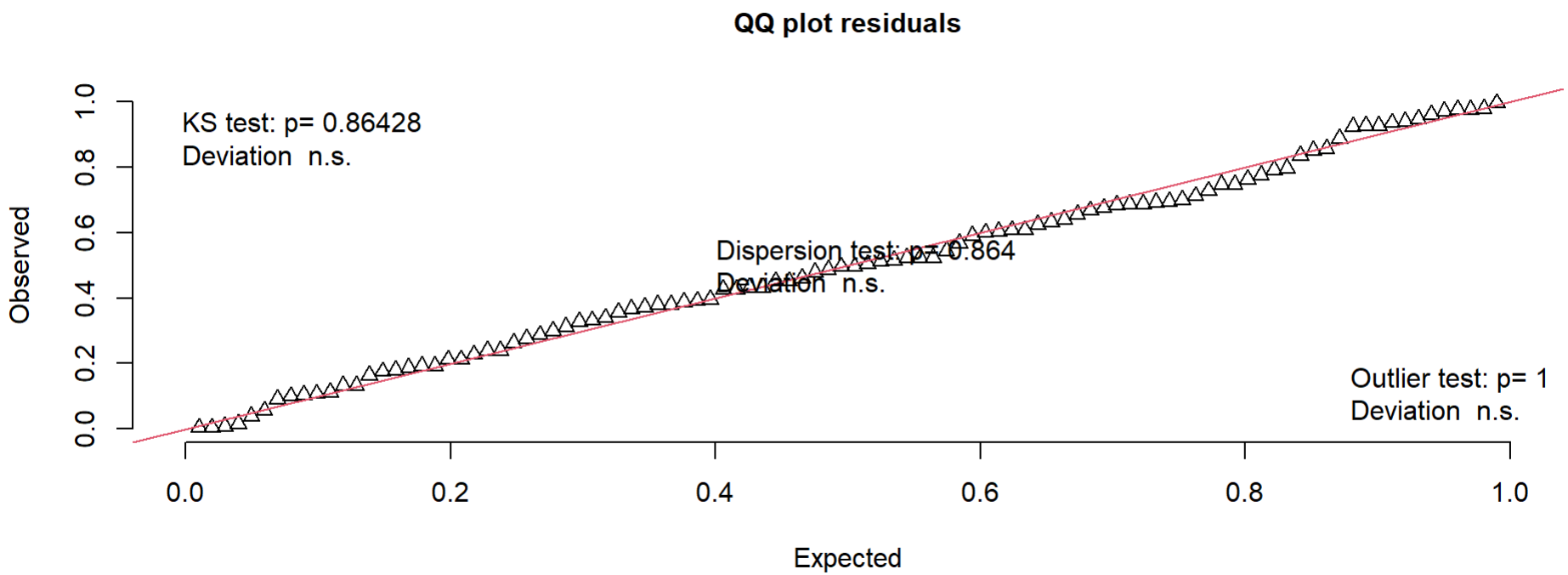
# 2. Independence

- If we consider that 10 males were sampled from each of 4 field sites, and account for this non-independence in our model, the relationship between male body size and number of mates is negative.



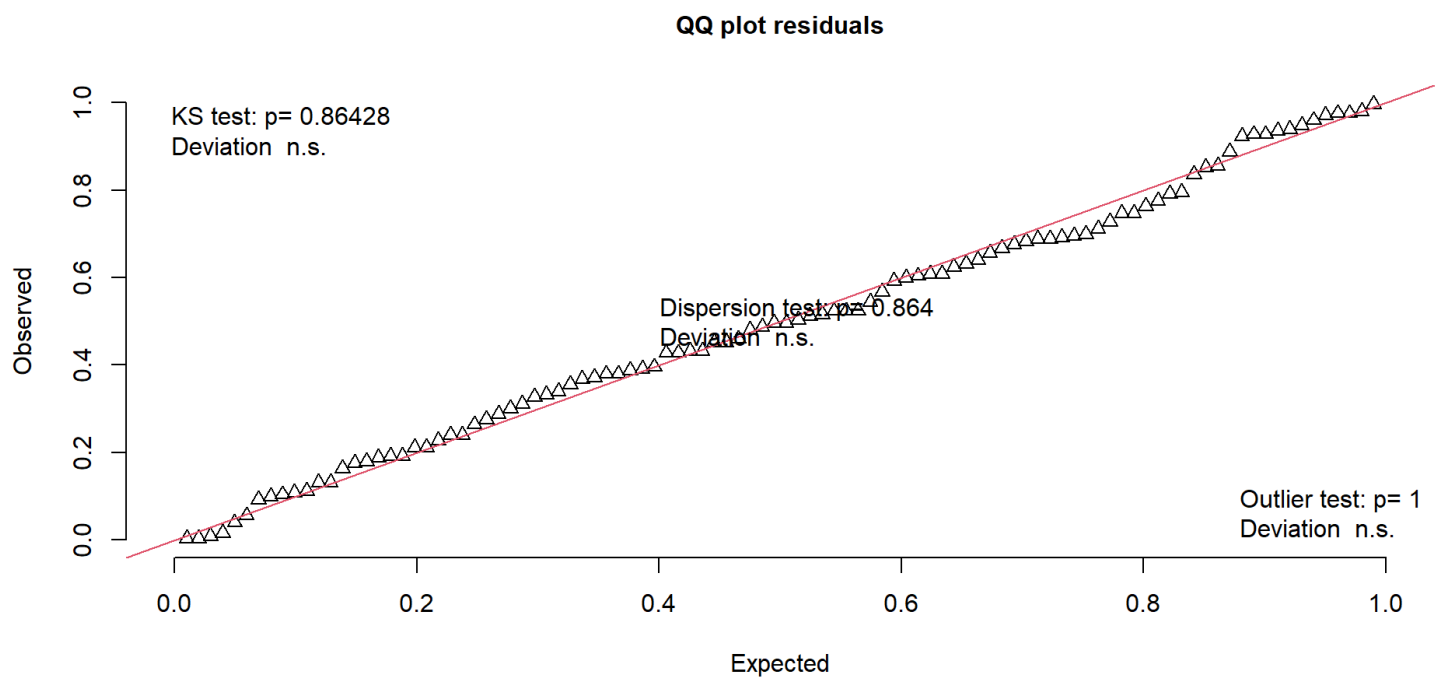
# 3. Normality

- Need to confirm that our responses are normally distributed for each level of `X`
  - Use a `Quantile-quantile (QQ) plot`
  - Residuals should cluster around slope = 1 curve, no pattern evident



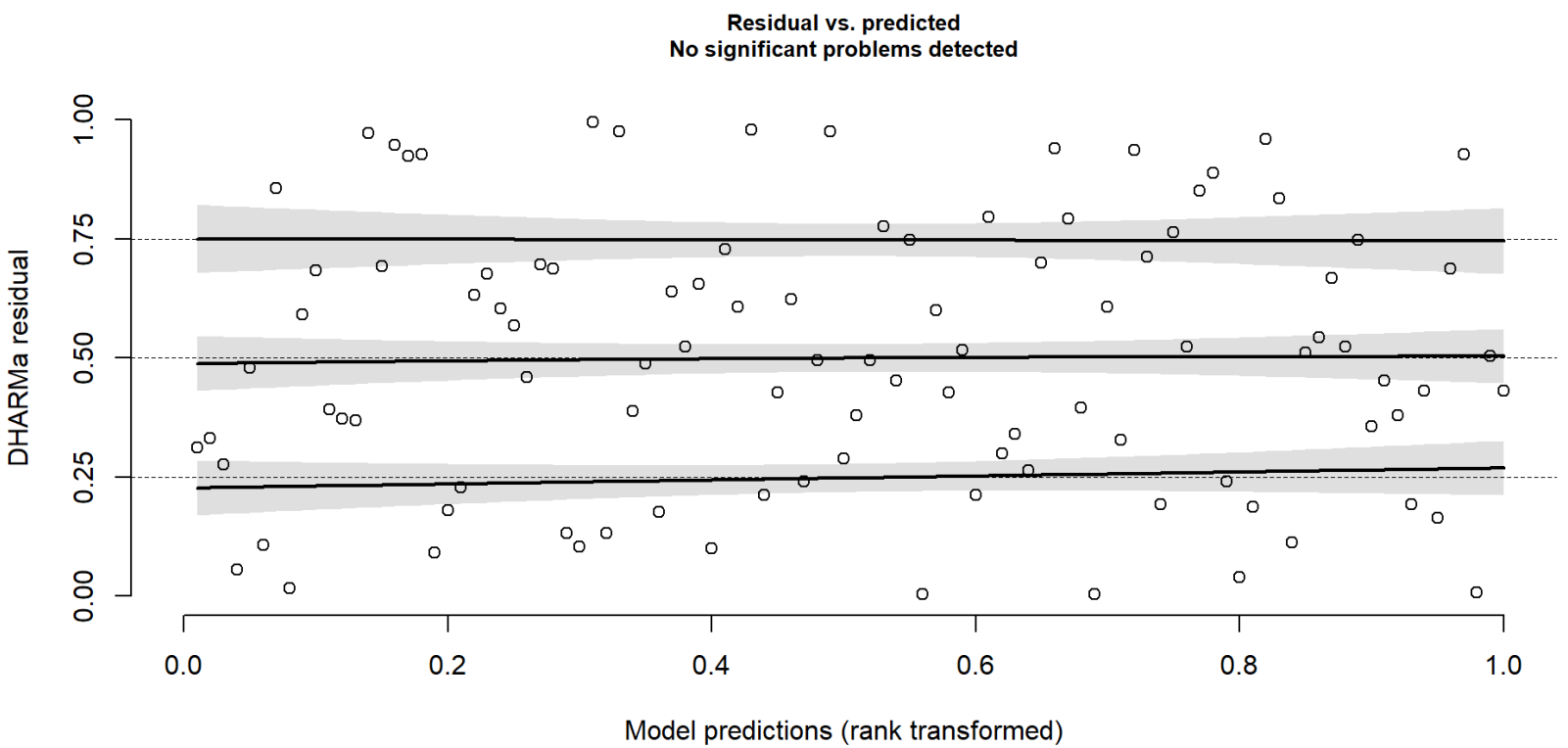
# 3. Normality

- Kolmogoroz-Smirnoff (KS) test is a formal test of whether residuals are significantly different from expectation under normal distribution
  - $P > 0.05$ = Model residuals are approximately normal
  - $P < 0.05$ = Model residuals are significantly different from normal



# 4. Equal variance

- The variation in responses are equal for each level in X
  - Spread of Y values on the y-axis should be similar for all values/groups of X
  - Bold lines should fall along the dotted y = [0.25, 0.50, 0.75] lines



# Model inference

# Model inference

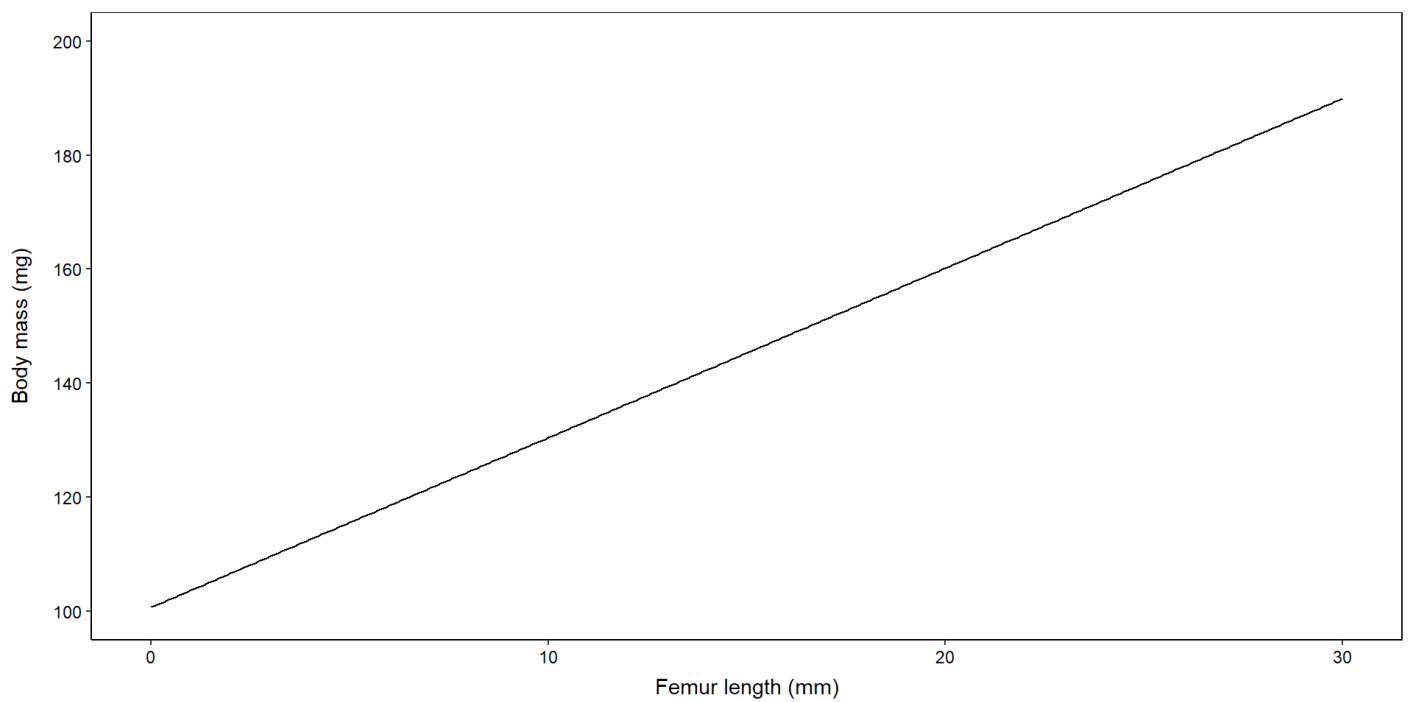
- We have confirmed that our statistical model was a good fit to the data, and we have even confirmed that `femurLength` was positively correlated with `bodyMass`.
  - But, how do we know if this result is *statistically significant* and calculate *p-values*, and all the things reviewers (and supervisors want!)?
  - Hint: We need to fit another model

# Null hypothesis significance testing (NHST)

- Most model inference in ecology is based on *Null hypothesis significance testing (NHST)*
  - The effect of a factor/covariate/predictor (e.g. `X` > `femurLength`) is evaluated against the hypothesis that there is no effect or relationship between the variable and `Y` (e.g. `bodyMass`)
    - To assess this, we need:
      1. A Global model, and
      2. A null model

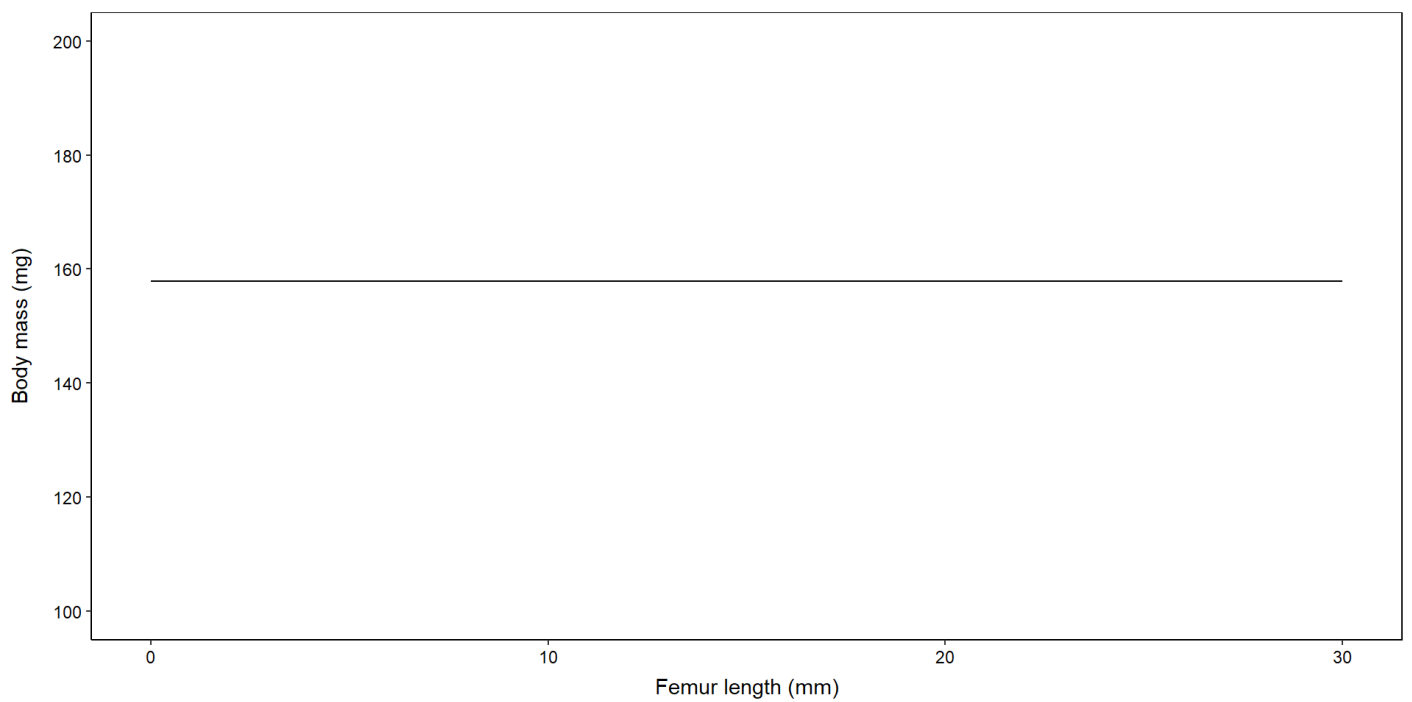
# What is a global model?

- Global model represents the alternative hypothesis ($H_1$)
  - The alternative hypothesis is that there is evidence for a statistically significant effect/relationship of our predictor variable(s) (`femurLength`) on our response variable (`bodyMass`)



# What is a null model?

- Null model represents the null hypothesis ($H_0$).
  - The null hypothesis is that there is no evidence for a statistically significant effect/relationship of our predictor variable(s) (`femurLength`) on our response variable (`bodyMass`)



# Null vs global model

1. Global model ($H_1$): There is a relationship between `femurLength` and `bodyMass`.

```
mod_global <- glm(
  data = df,
  family = gaussian(link = "identity"),
  bodyMass ~ 1 + femurLength
)
```

2. Null model ($H_0$): There is no statistical evidence for a relationship between `femurLength` and `bodyMass`.

```
mod_null <- glm(
  data = df,
  family = gaussian(link = "identity"),
  bodyMass ~ 1
)
```

# Null vs global model

# Hypothesis testing

- Finally, we have to actually perform a hypothesis test
  - Was the global model ($H_1$) or the null model ($H_0$) better supported by the data?
- To do this, we use the **Likelihood Ratio Test (LRT)**
  - In R, we perform the LRT using the following code:
  - `lmtest::lrtest(mod_null, mod_global)`
  - Asks which model likely fits the data better (goodness-of-fit test)

# Likelihood Ratio Test (LRT)

- This test gives us our test statistic (X2), degrees of freedom (df), and our sacred p-value.

```r
# Perform LRT
lmtest::lrtest(
  mod_null,
  mod_global
)
```

```
Likelihood ratio test

Model 1: bodyMass ~ 1
Model 2: bodyMass ~ 1 + femurLength
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -417.40
2   3 -305.47  1 223.86  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Reporting LRT

```r
# Perform LRT
lmtest::lrtest(
  mod_null,
  mod_global
)
```

```
Likelihood ratio test

Model 1: bodyMass ~ 1
Model 2: bodyMass ~ 1 + femurLength
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -417.40
2   3 -305.47  1 223.86  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There is a statistically significant relationship between `femurLength` and `bodyMass` (X2 = 223.86, df = 1, $P < 0.001$).
  - We know that one of the models was better than the other because the $P < 0.05$
  - We then can tell that model 2 (`mod_global`) is the better model because it has a **higher** likelihood (`logLik = -305.47`) than model 1 (`mod_null`) (`logLik = -417.40`)

# Plotting model predictions (marginal effects)

# Marginal effects

Easiest way to present results is typically a *marginal effects plot*.

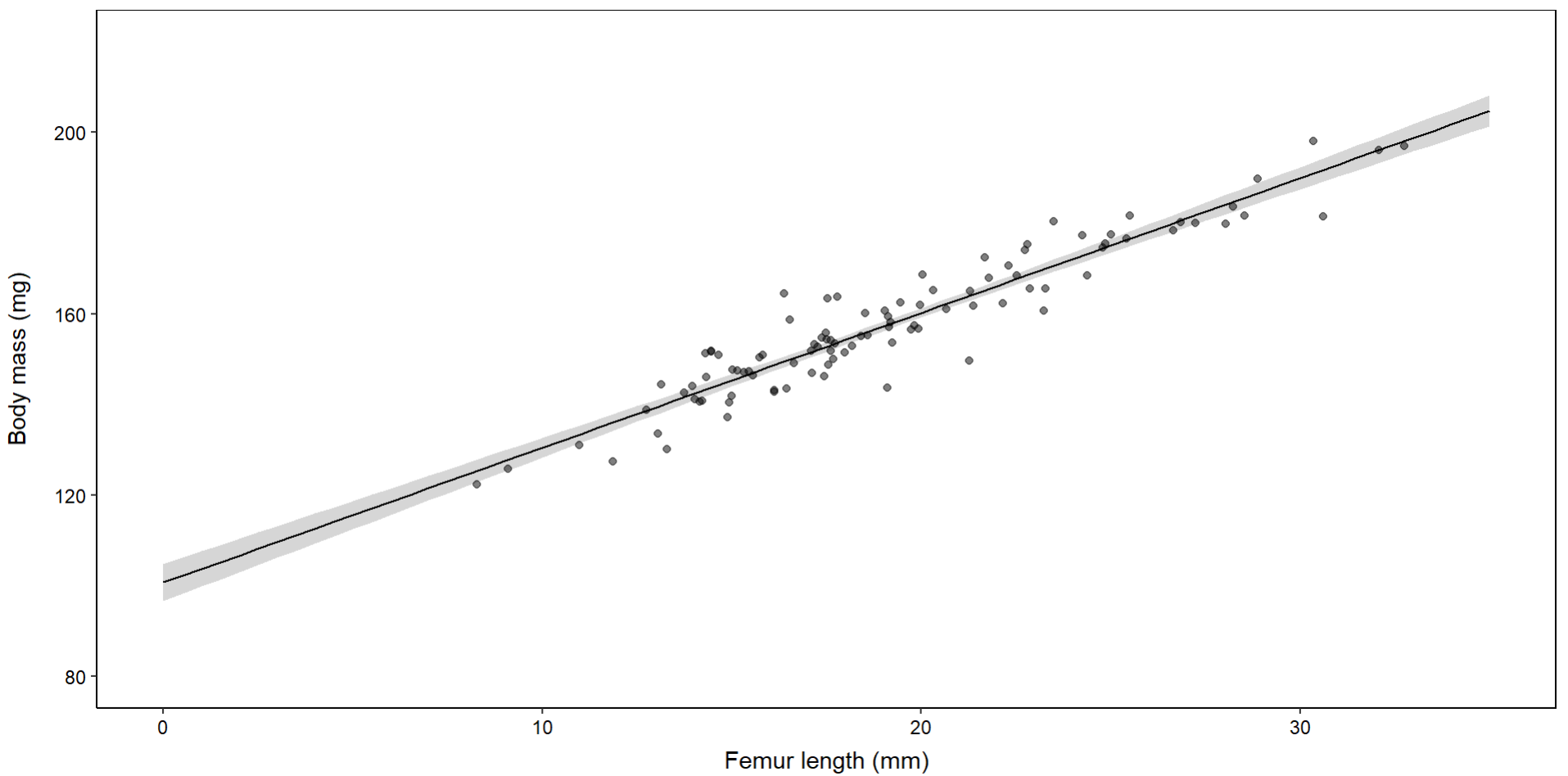
- Marginal effects show the relationship between our predictor(s) and response variable, holding all other predictors in the model constant or at a specified value
  - In this example, we only have 1 predictor, so the marginal effects plot simply shows the relationship between `femurLength` and `bodyMass`

# Extracting marginal means

```r
# Extract expected relationship between X and Y
preds <- ggeffects::ggeffect(
  model = mod_global,
  terms = c("femurLength [0:35 by = 0.5]"),
  type = "fixed",
  interval = "confidence"
  ) %>%
  # Convert predictions into a data.frame
  as.data.frame() %>%
  # Rename columns for easier plotting
  dplyr::mutate(
    femurLength = x
  )
```

# Plot marginal effect plot

Plot | Code



# Reporting your results

There was a statistically significant relationship between `femurLength` and `bodyMass` (X2 = 223.86, df = 1, $P < 0.001$). The beta-coefficient for this relationship was 2.97, indicating that for every 1mm increase in `femurLength`, `bodyMass` increases by 2.97mg, on average. (Insert your plot).