Assumptions of Statistical Tests

R Club

Jacinta Kong

16/3/2022

Partitioning variation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Explain as much variation in Y as possible using the fewest terms possible (β)

- β partitions variation to each potential *fixed* source of variation
  - Predictor variables X_1 & X_2
  - Interactions $X_1 X_2$
- Any *random* effects (not shown here, γ)

Residual error ε

Random *or* ***residual error*** $=$ *unexplained variation*

- All frequentist tests make assumptions about ε - LM, GLM etc
- As ε is *random*, assumptions also apply to Y

4 assumptions

1. Normality
2. Heterogeneity of variance
3. Independence
4. Fixed X

- Linearity

2 more for "traditional" ANCOVA

5. Covariate values cover a similar range across groups
6. Regression slopes are similar across groups

Checking assumptions

Residual plots show relationship between residuals and model

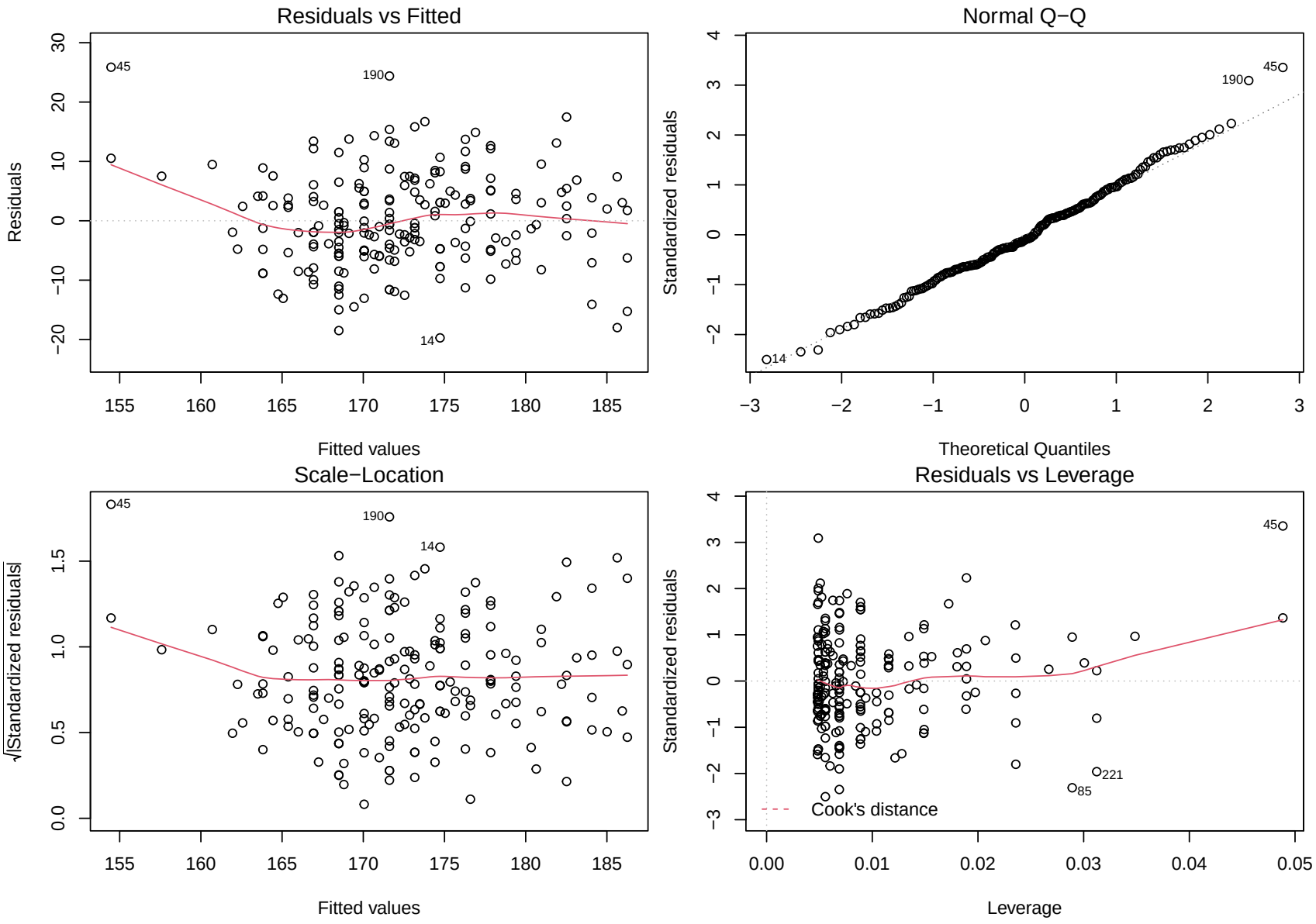
```
plot(lm(Y ~ X, data))
```

1. Residuals vs fitted values
2. Standardised residual quantile quantile plot
3. Standardised residuals vs fitted values
4. Residuals vs Leverage

- Standardised residual = residual / standard deviation
  - Control for unequal variance

Data from `MASS`

Residual plots



1. Normality

Population Y values and error terms (ε) are normally distributed for each level of the predictor variable (X)

- Data follows normal distribution
- Doesn't apply to non-Gaussian GLM
- Check:
  - Histogram of Y
  - Quantile-Quantile plot of Y and ε

Histograms

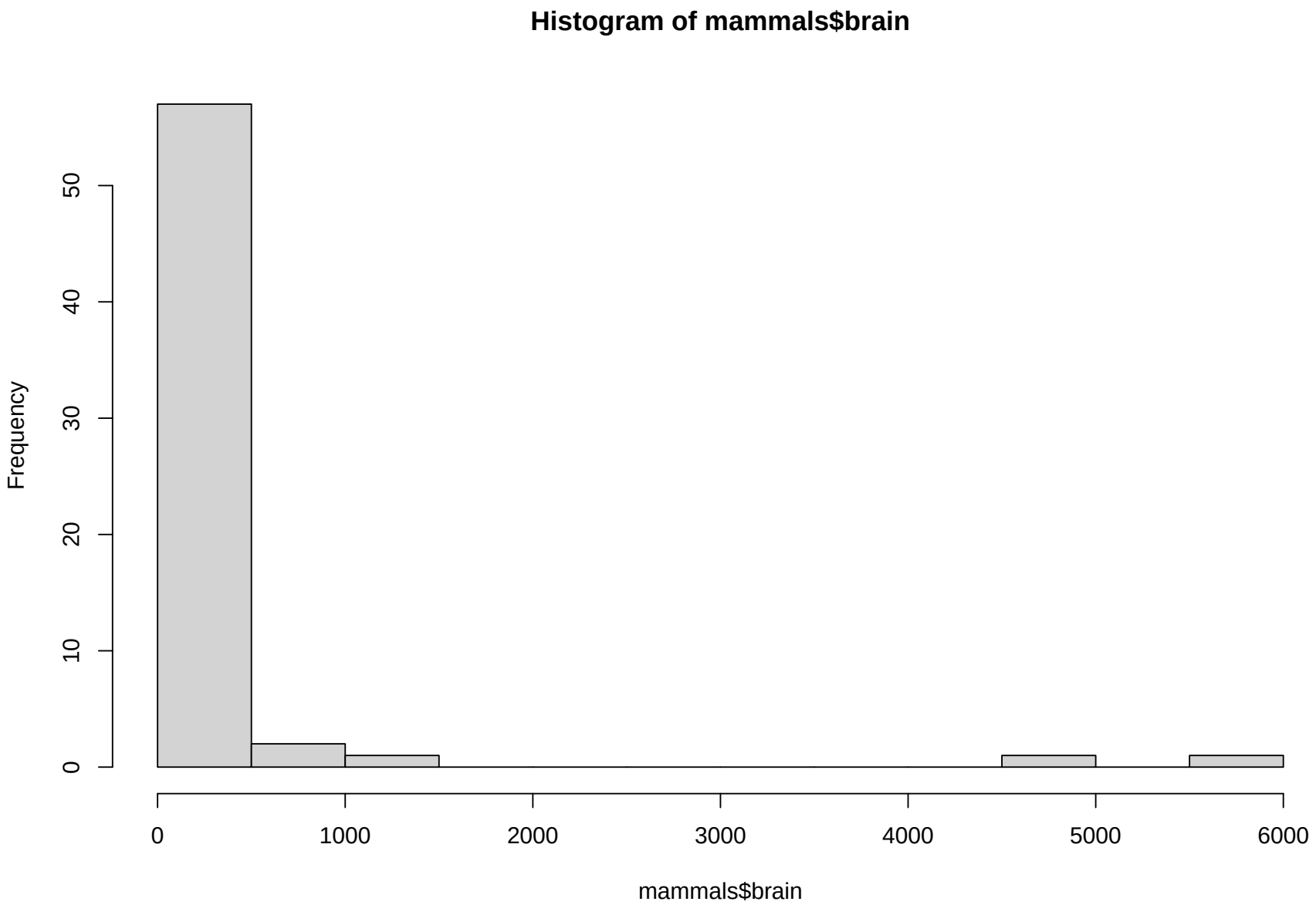


Figure 1: Right skewed mammal brain size

Quantile-quantile plot

Plots theoretical quantiles of a normal distribution against observed quantiles

```
qqnorm(data$Y)
```

- 1:1 relationship if normal
- Deviation indicates skewedness

Add theoretical line to `qqnorm`:

```
qqline(data$Y)
```

Mammal brains

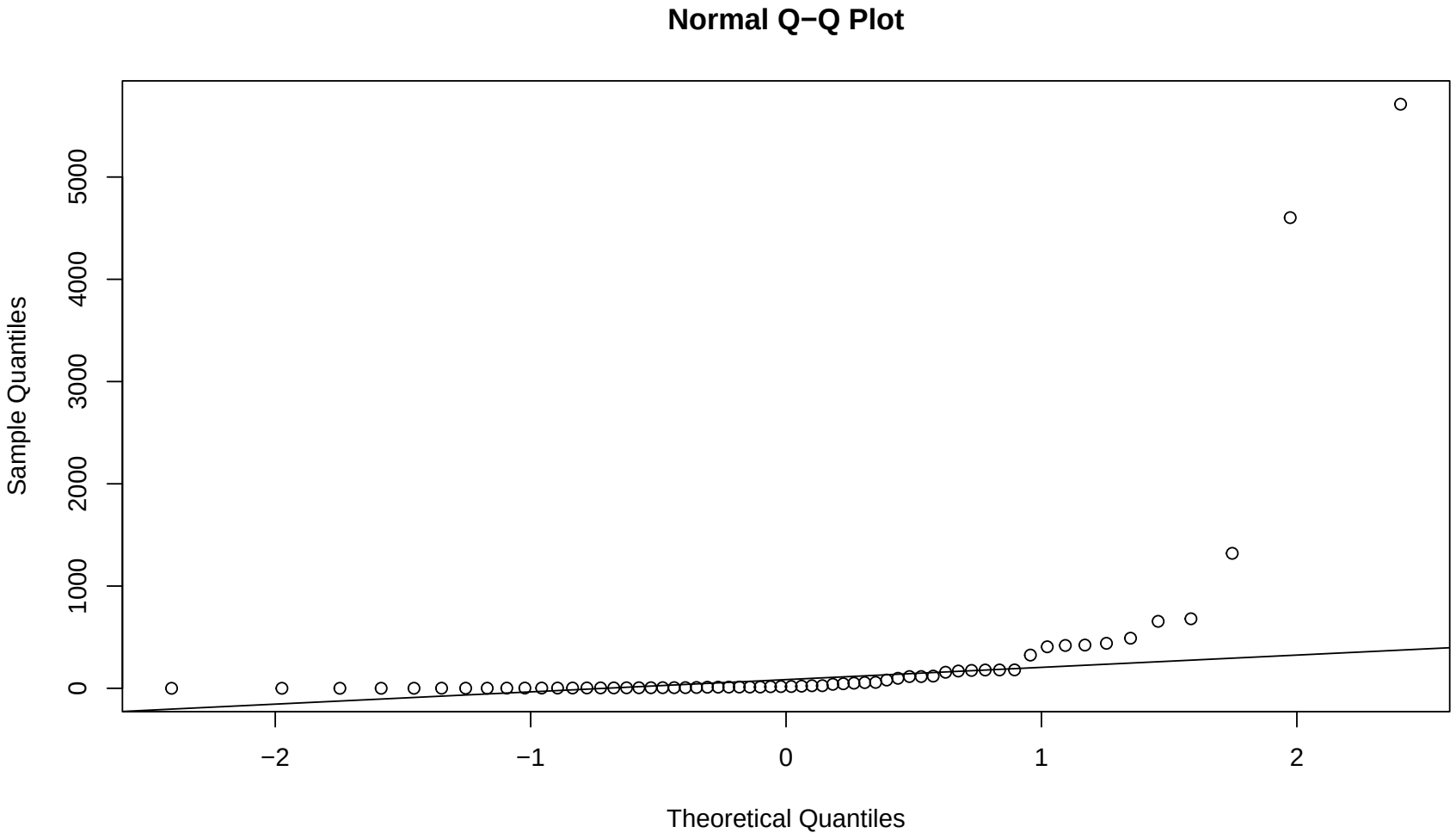
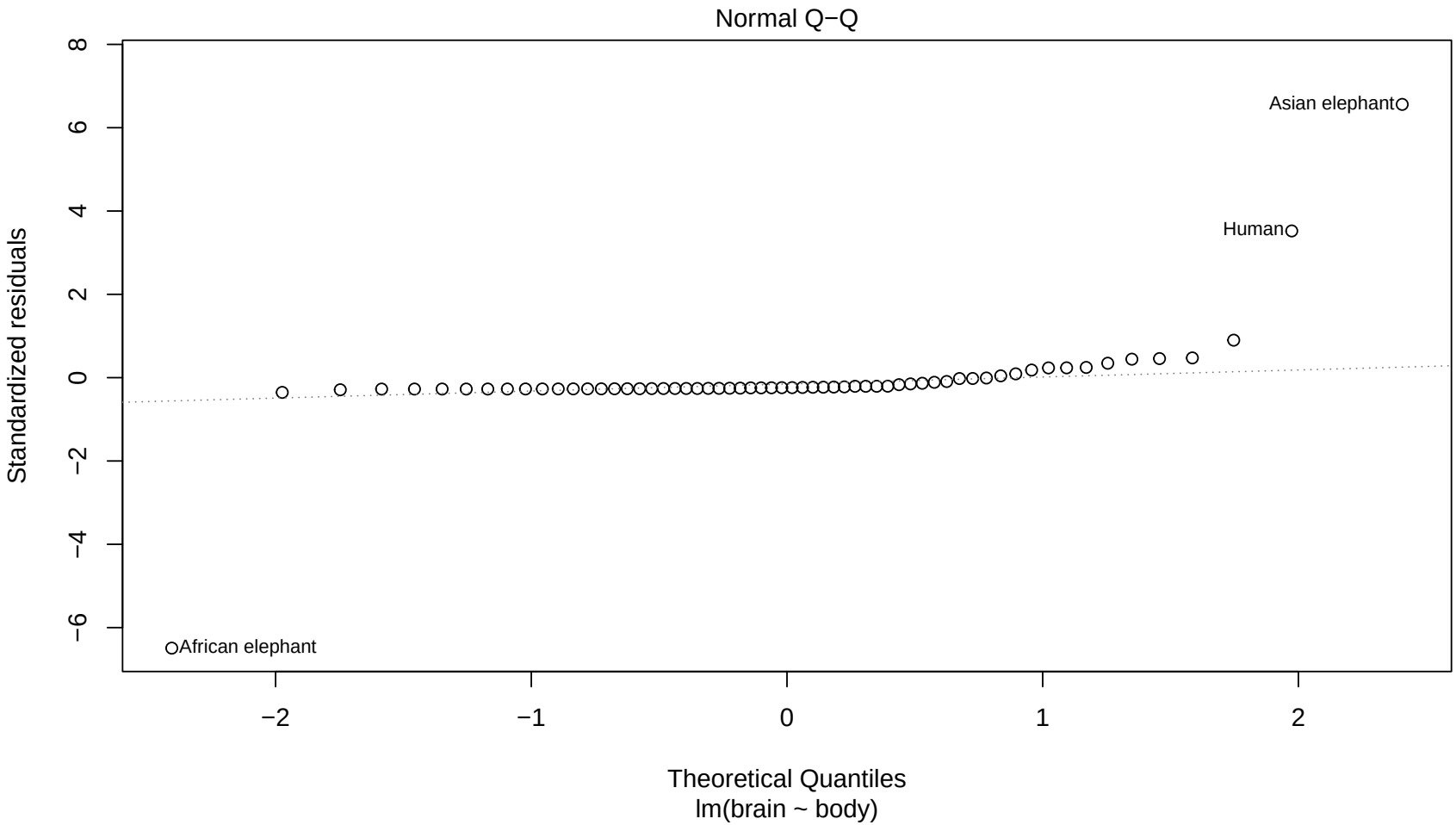


Figure 2: Quantile-Quantile plot of mammal brains

Model residuals

```
mammal_brains <- lm(brain ~ body, mammals)
plot(mammal_brains, which=c(2))
```



Non-normality

- Small sample sizes - Central Limit Theorem

Non-normality

- Small sample sizes - Central Limit Theorem
- Ignore it - Robust to some skewedness

Non-normality

- Small sample sizes - Central Limit Theorem
- Ignore it - Robust to some skewedness
- Use alternative tests

Non-normality

- Small sample sizes - Central Limit Theorem
- Ignore it - Robust to some skewedness
- Use alternative tests
  - GLM

Non-normality

- Small sample sizes - Central Limit Theorem
- Ignore it - Robust to some skewedness
- Use alternative tests
  - GLM
  - Non-linear regression

Non-normality

- Small sample sizes - Central Limit Theorem
- Ignore it - Robust to some skewedness
- Use alternative tests
  - GLM
  - Non-linear regression
  - Non-parametric test

Non-normality

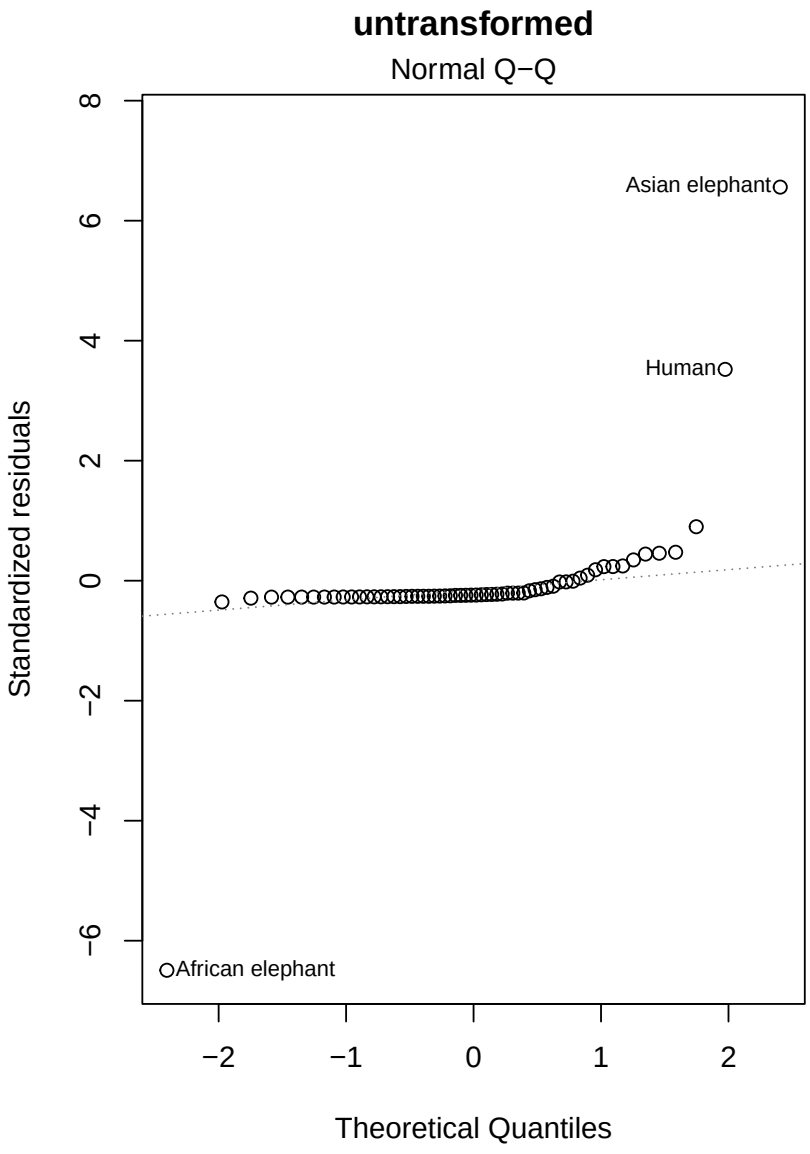
- Small sample sizes - Central Limit Theorem
- Ignore it - Robust to some skewedness
- Use alternative tests
  - GLM
  - Non-linear regression
  - Non-parametric test
- Transformation

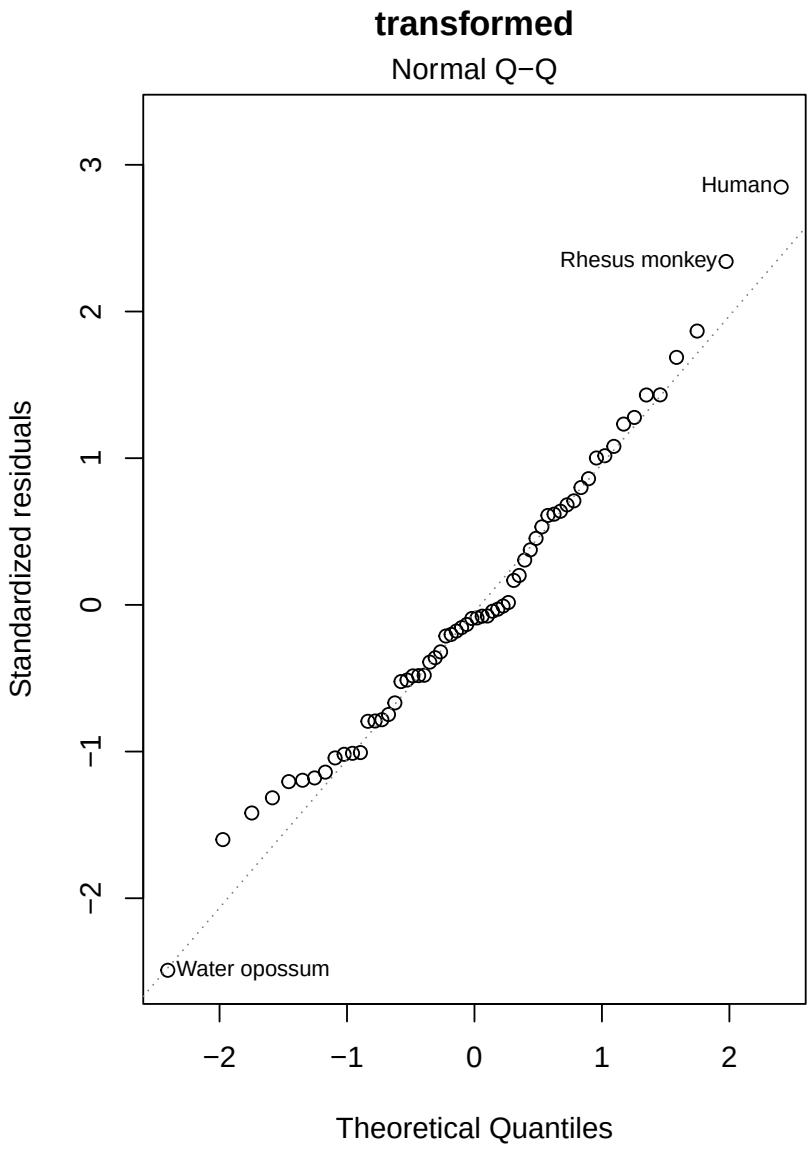
Transformations on Y

Spread out Y more evenly

- $\log_{10}$ or natural log - positive non-0 numbers
- square root, cube root - positive including 0
- inverse

$\log_{10}$ mammal brains





2. Homogeneity of Variance

Population Y values and error terms (ε) have the same variance for each level of the predictor variable (X)

- Also called homoscedasticity
- Variances are the same - important for Analysis of Variance!

Check variances and residuals:

- Quantile plot
- Relationship with fitted values (predictions of Y from model)

Examples: Uneven standard deviation

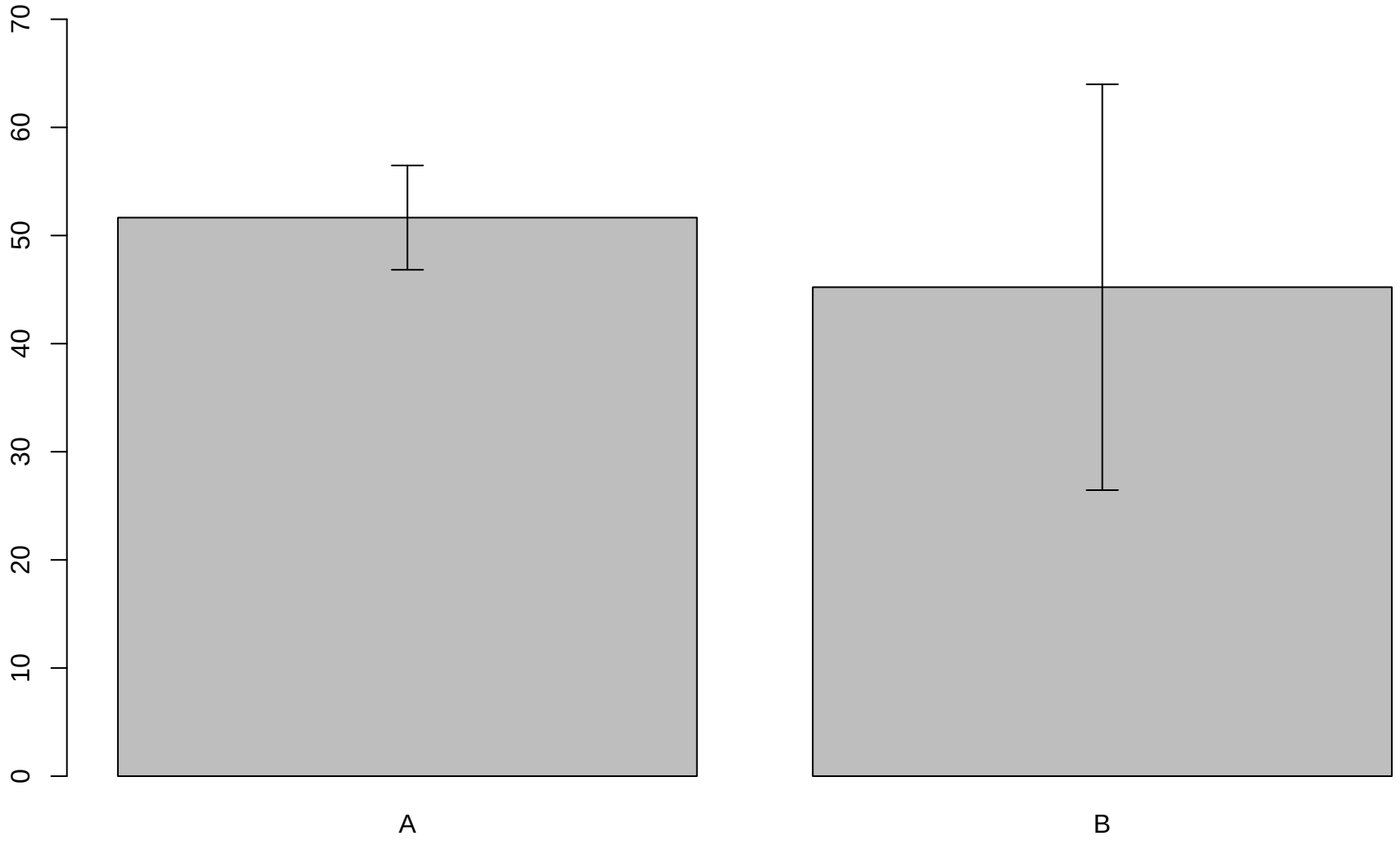


Figure 3: Bar plot of mean of two groups (A and B). Error bars indicate standard deviation

Examples: Non-independence in Y

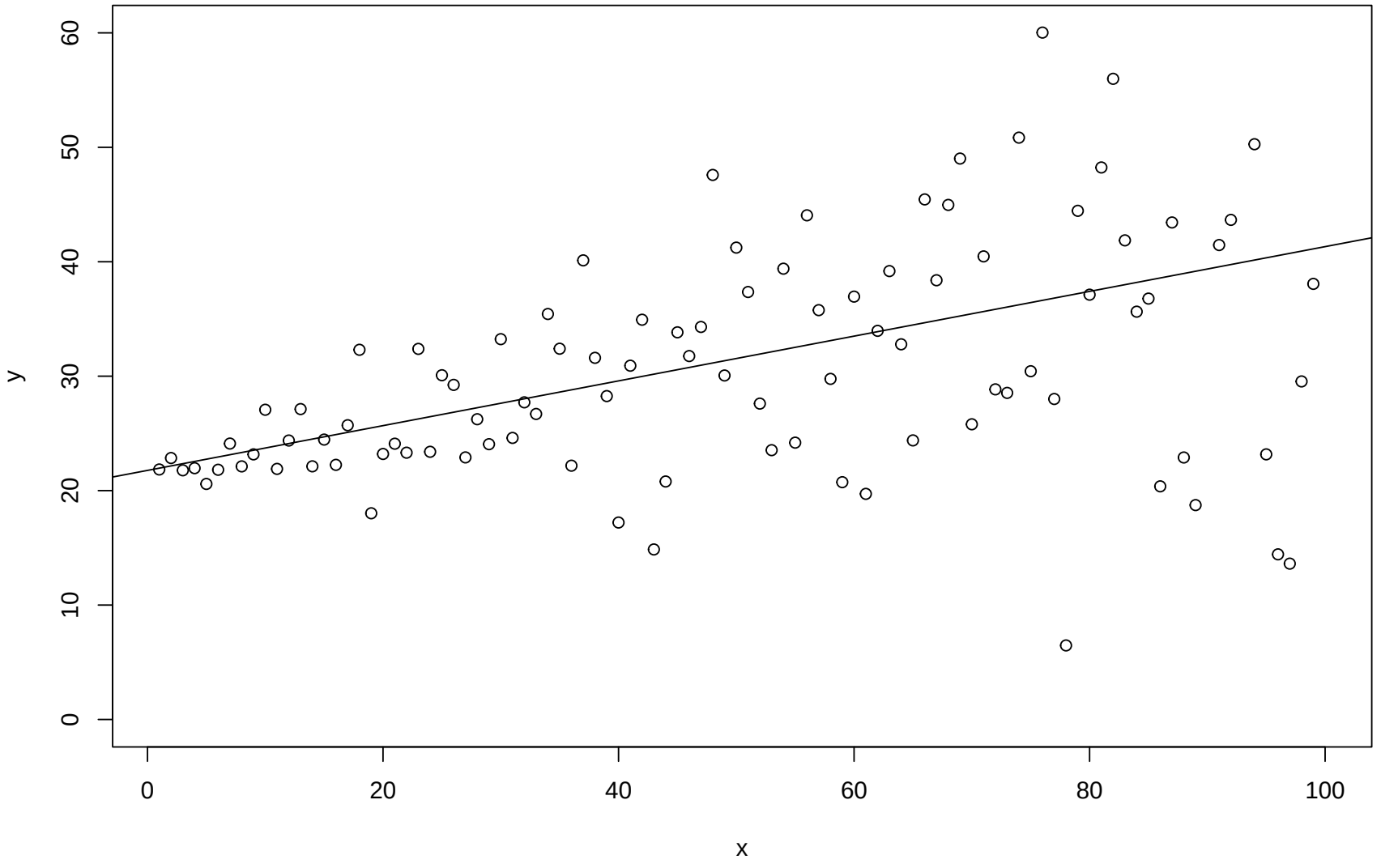


Figure 4: A scatter plot and a fitted model

Examples: Residual plots

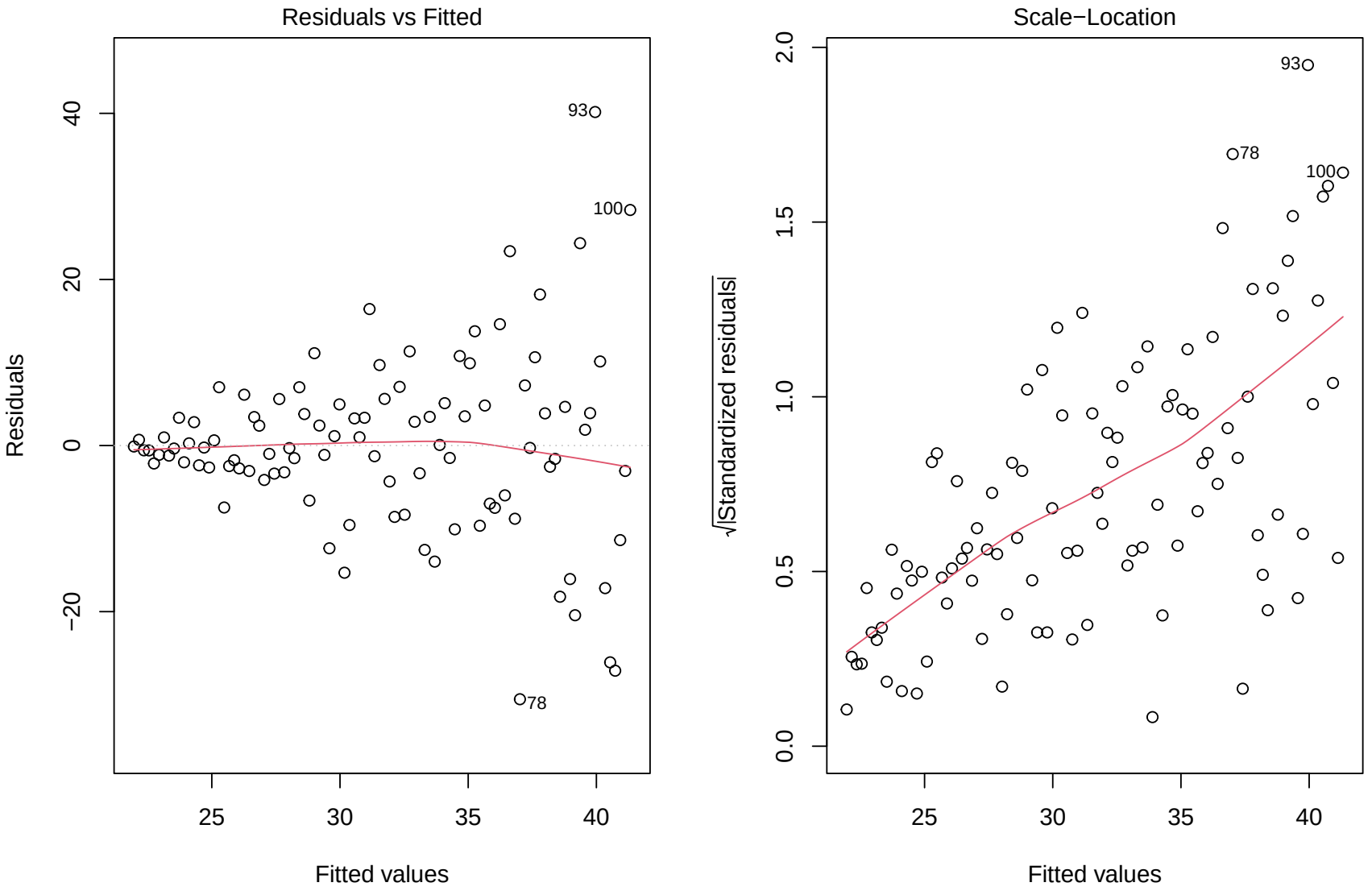


Figure 5: Shotgun pattern

Dealing with heteroscedasticity

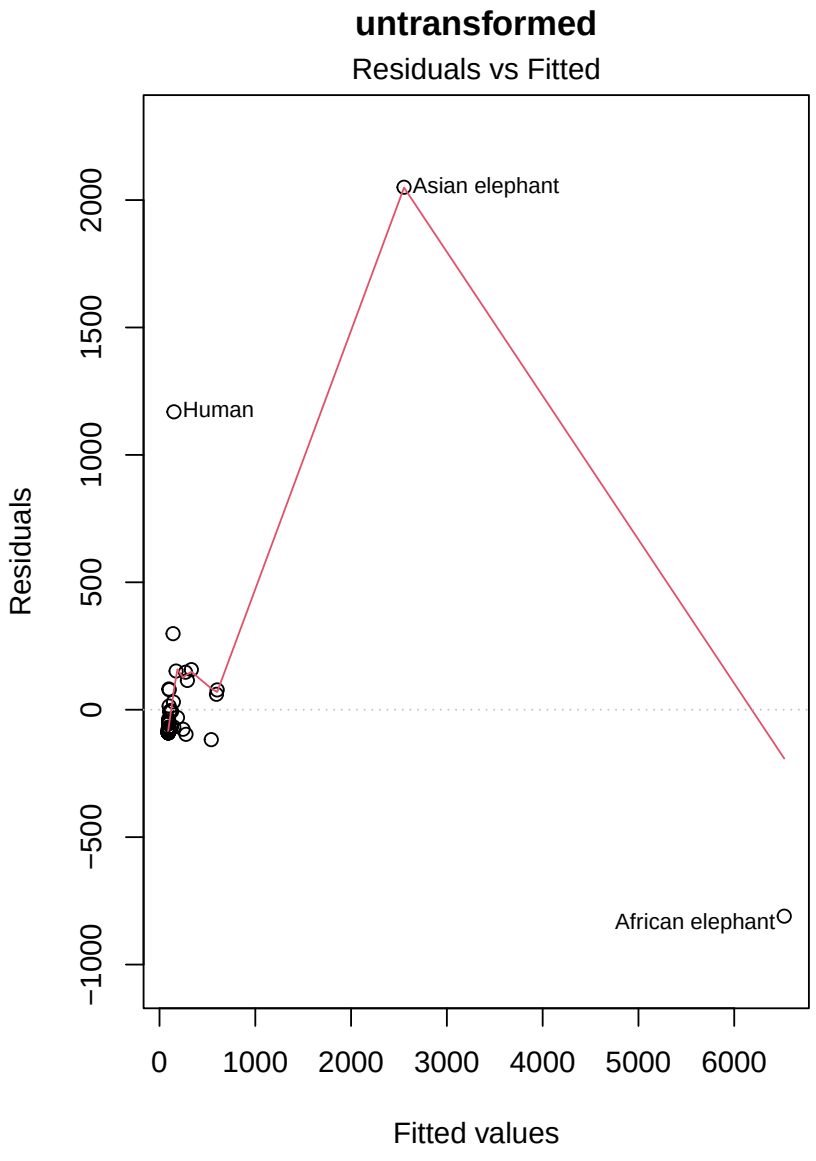
Causes:

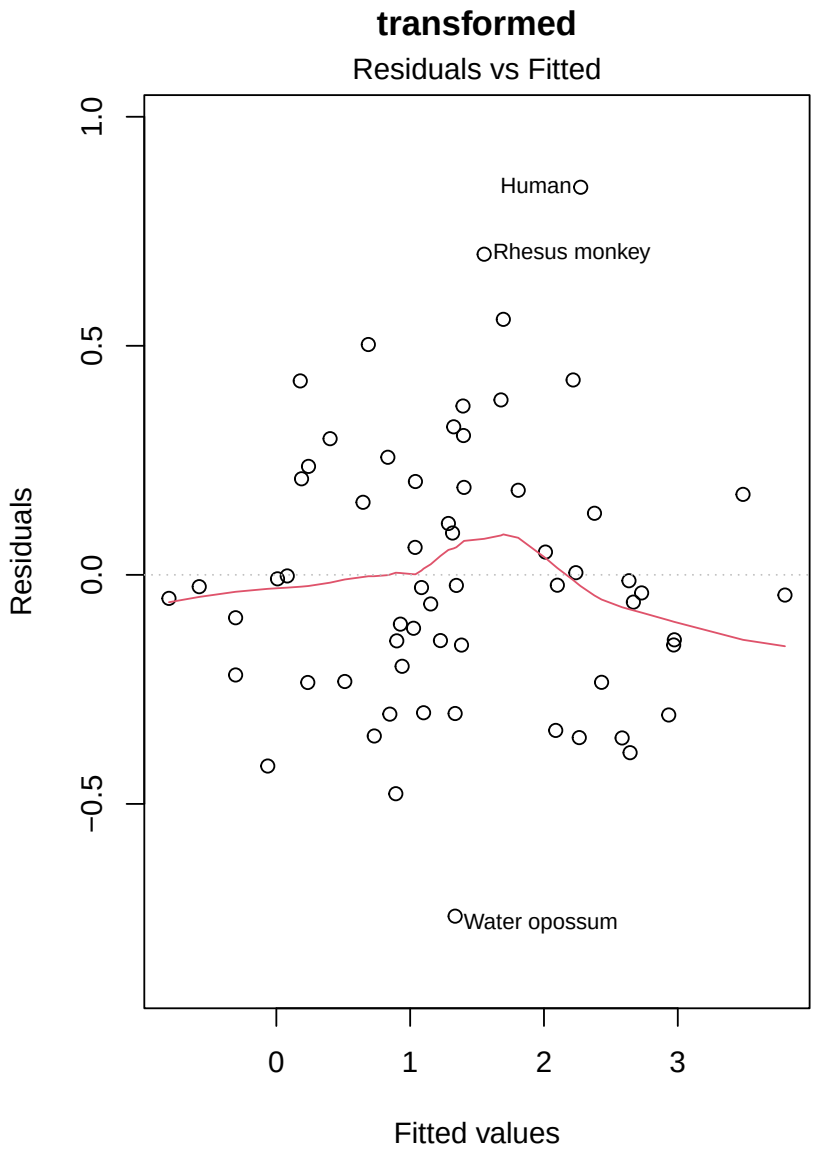
- Small sample size
- Outliers
- Non-normal distribution
- Non-independent values (e.g. time series)

Solutions:

- Balanced experiments
- Sufficient sample size
- As with normality - transformation
- Advanced linear regression methods

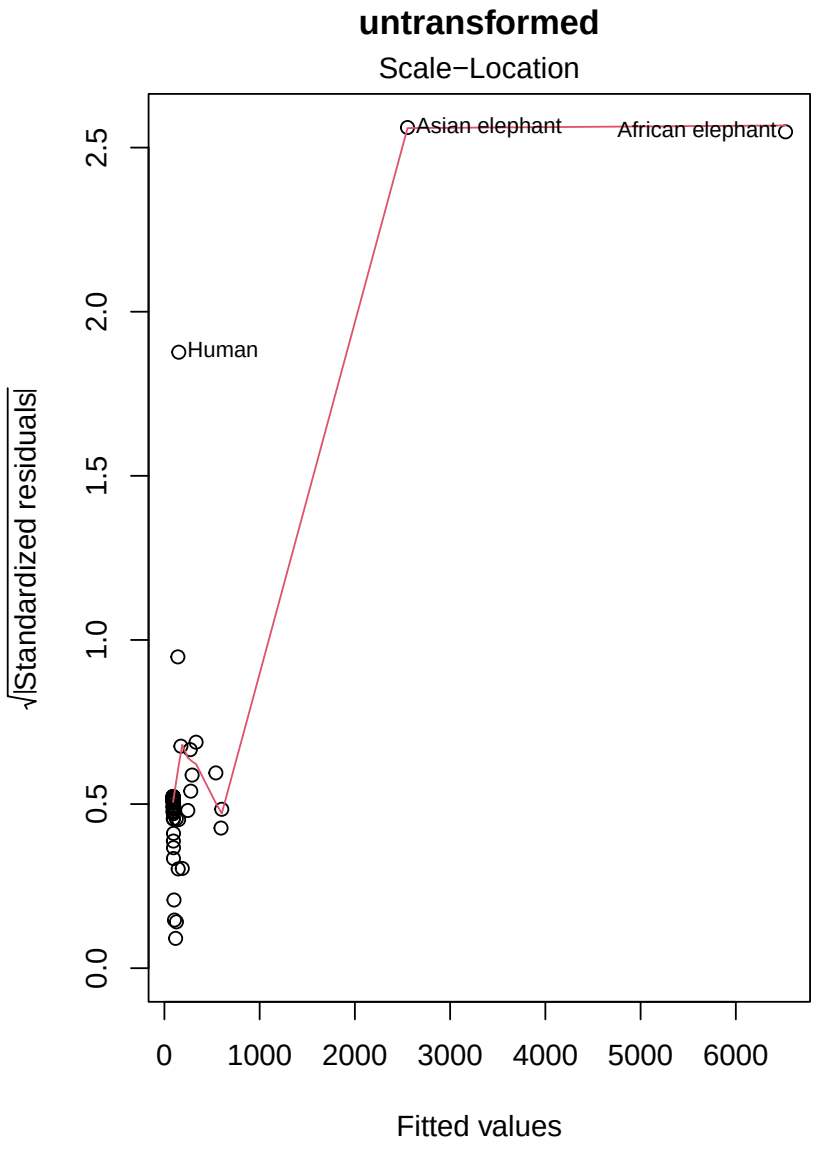
Example mammals: Residuals

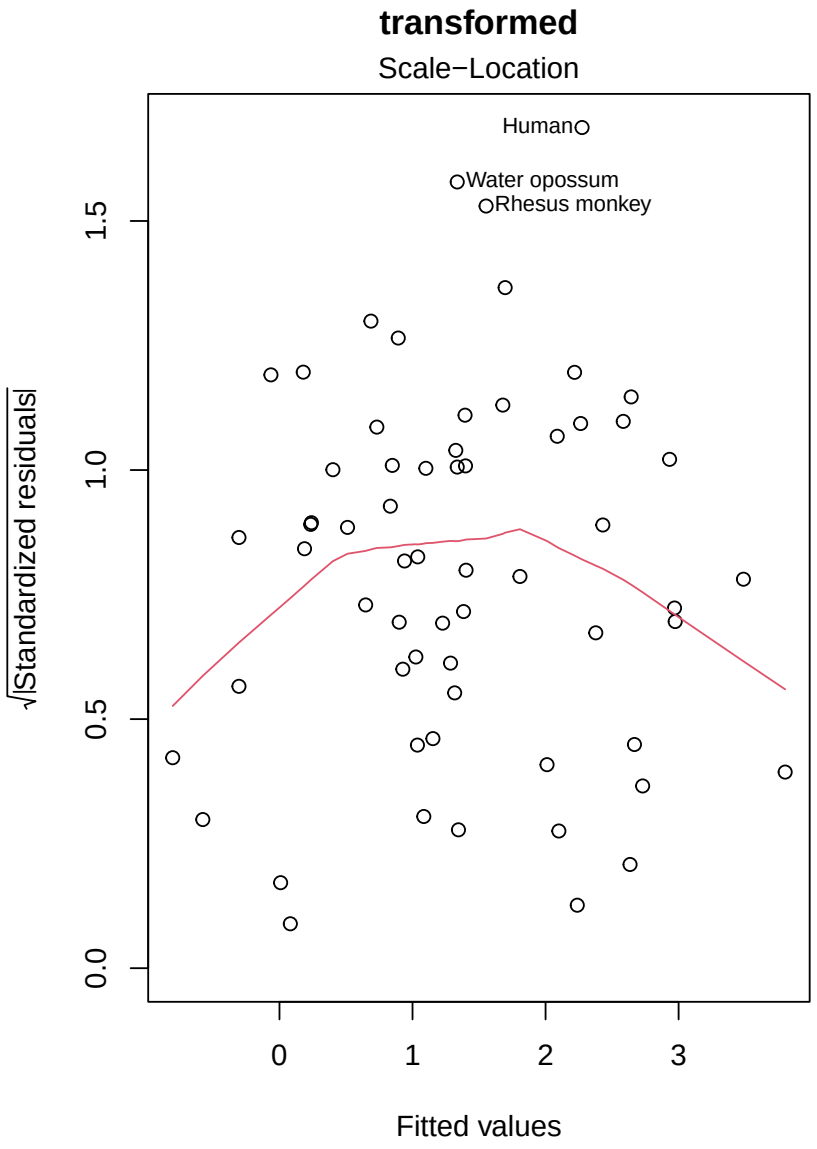




Example mammals: Standardised residuals

Also quantile quantile plot of standardised residuals.





3. Independence

Population Y and error terms (ε) are independent

- Autocorrelation
- Effect of experimental design
  - Time series
  - Pseudo-replication
  - Repeated measures
- Important for GLM
- Check Residuals vs X values or row number

Solutions

- Random effects model
- Drop variables
- Careful experimental design
- Advanced analyses for repeated measures (e.g. paired t-test, repeated measures ANOVA)

4. Fixed X

The predictor variable is fixed - a known constant, can explain all variation

- Type I model - often broken in biostats
- Type II model - random effects
- Type III model - mixed effects

Changes F ratio in ANOVA.
Use more advanced estimation functions, e.g. `lmer`, `nlme` and (restricted) maximum likelihood.

Outliers

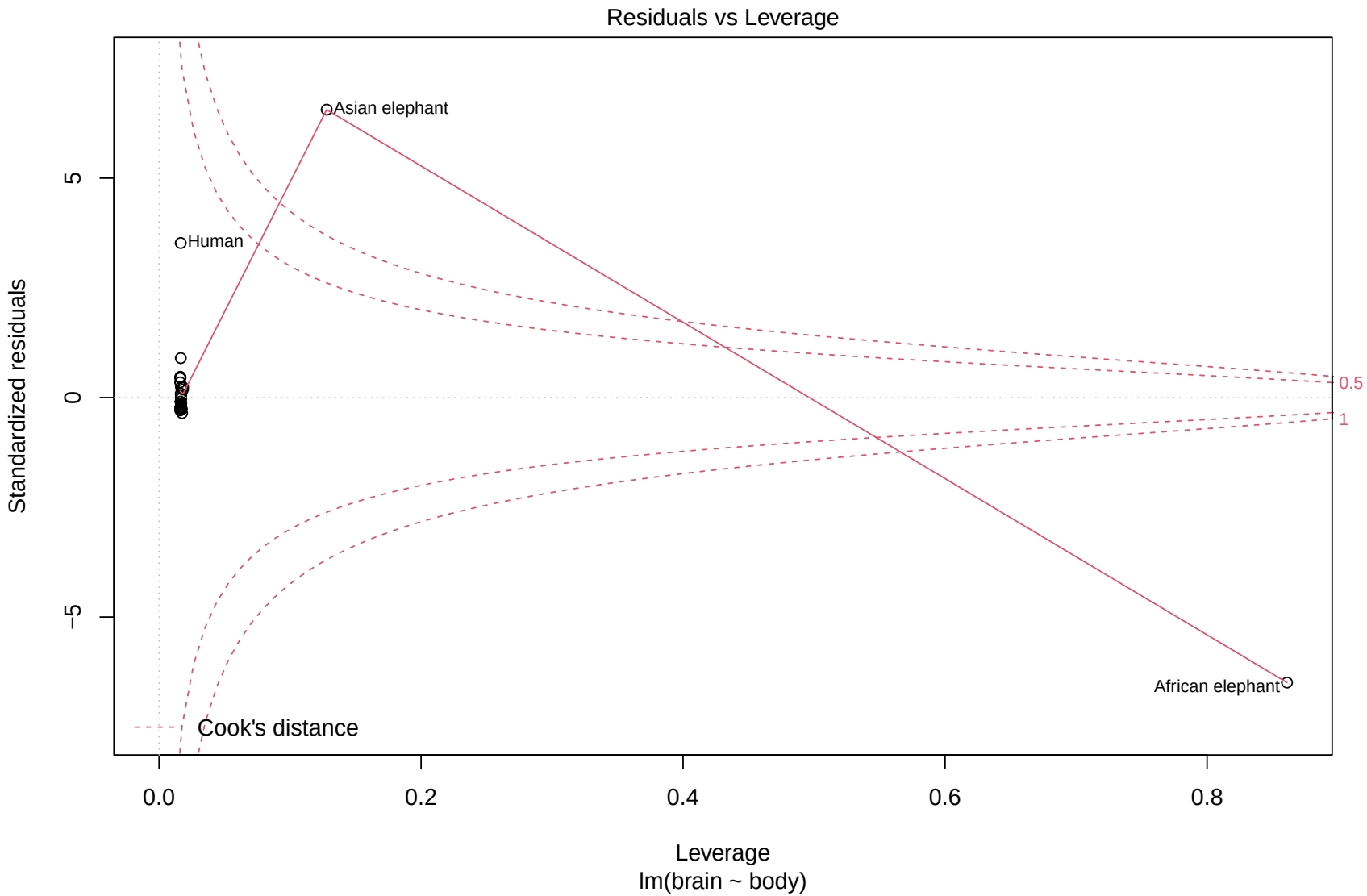
- Assess before fitting - e.g. 1.5 IQR
- Evaluate wrt biological context
- Leverage = how much X influences Y
- Influence = how much X influences slope of line (Cook's Distance)

Other residual plots

- Plot 4: Cook's Distance vs observation number
- Plot 6: Cook's Distance vs Leverage

Mammal outliers

```
plot(mammal_brains, which=c(5))
```



Summary

Check assumptions. Make sure stats is appropriate

- Plan stats from the start
- Formal tests of assumptions
- Bootstrapping
- Bayesian approaches