Continuous bivariate data

Presenter: Steve Baskauf
steve.baskauf@vanderbilt.edu

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

CodeGraf landing page

- vanderbi.lt/codegraf

Types of analyses

Jean & Alexander Heard
LIBRARIES

Categories of analyses

- Common characteristics: two variables, both continuous (numeric)

- Linear regression
  - as a model to predict one variable from another
  - as a statistical test for assessing significance of an effect
- Correlation
  - to assess strength of relationship between the variables

Decision tree

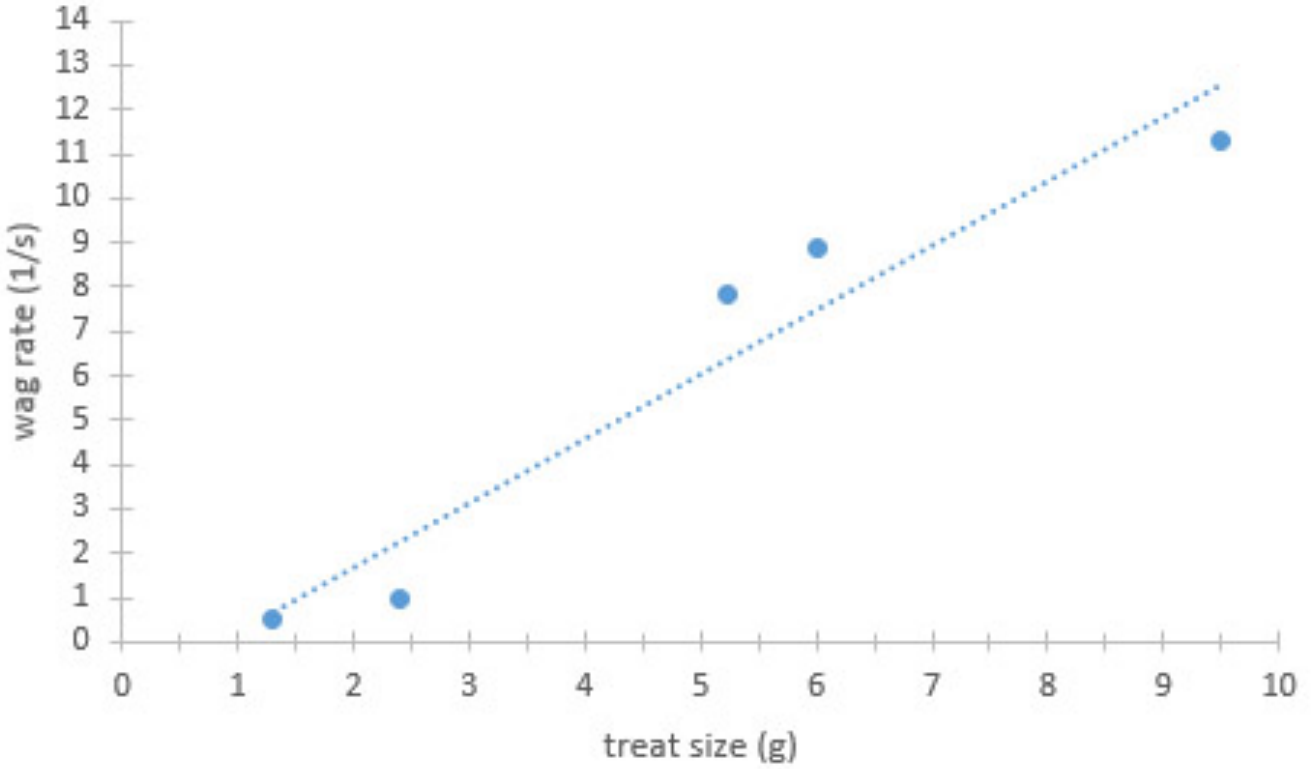
1. Am I assuming cause and effect?
   - Yes: linear regression as a test
   - No: go to 2
2. What do I want to know?
   - strength of the relationship: correlation
   - prediction: linear regression as a model for prediction

Linear regression for prediction

Jean & Alexander Heard
LIBRARIES

What is linear regression?

- Linear regression determines the straight line that best fits a set of points.
- The line minimizes the distance between it and each point (least squares method).



Linear regression in R

- Create a linear model. Both variables must be continuous (numeric)

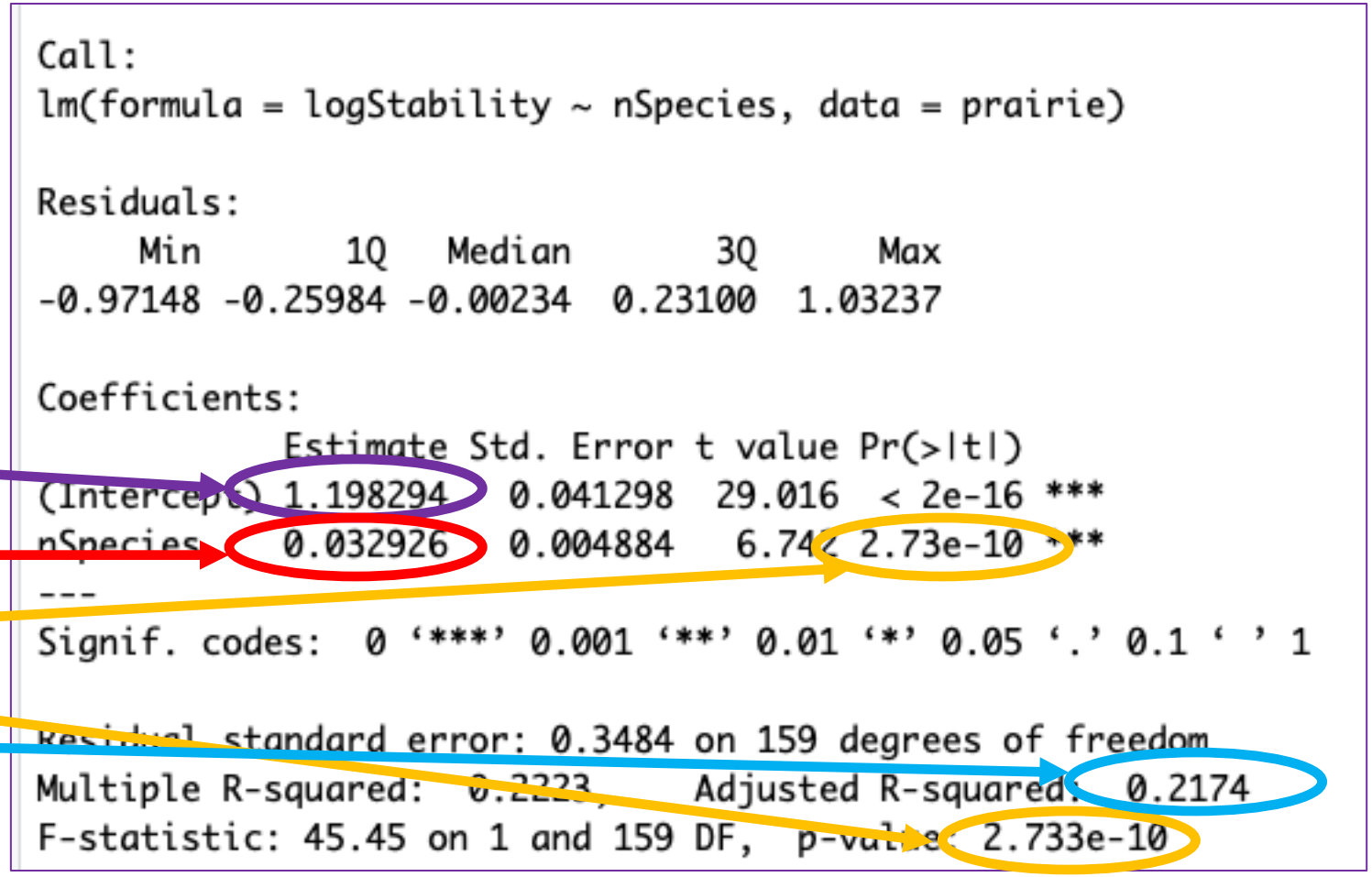
```
model <- lm(Y ~ X)
```

- Display the model summary

```
summary(model)
```

What information do we get from linear regression?

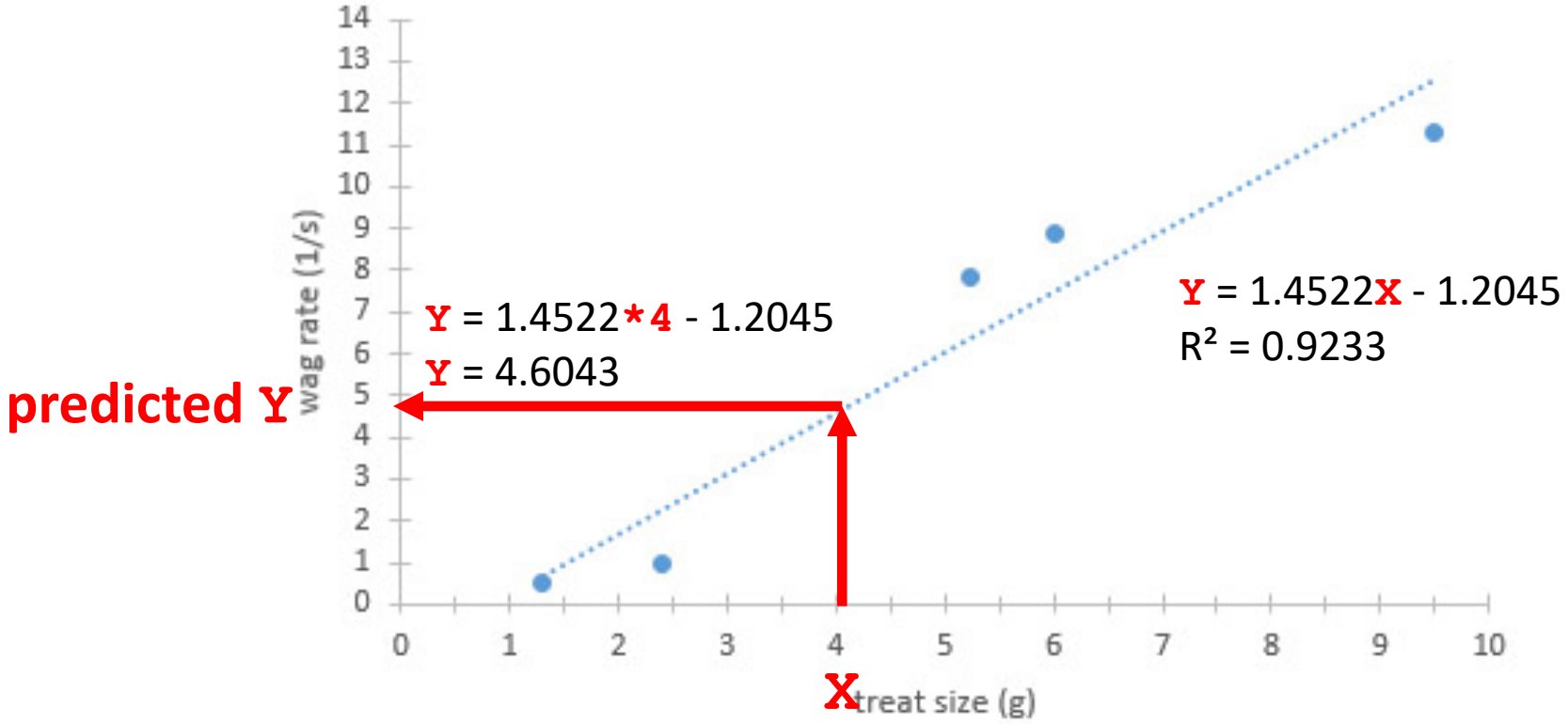
- Y intercept
- slope
- P value
- R squared (R^2)
- other stuff we don't care about very much



```
Call:
lm(formula = logStability ~ nSpecies, data = prairie)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97148 -0.25984 -0.00234  0.23100  1.03237

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.198294   0.041298  29.016  < 2e-16 ***
nSpecies    0.032926   0.004884   6.74< 2.73e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3484 on 159 degrees of freedom
Multiple R-squared:  0.2223,    Adjusted R-squared:  0.2174
F-statistic: 45.45 on 1 and 159 DF,  p-value: 2.733e-10
```

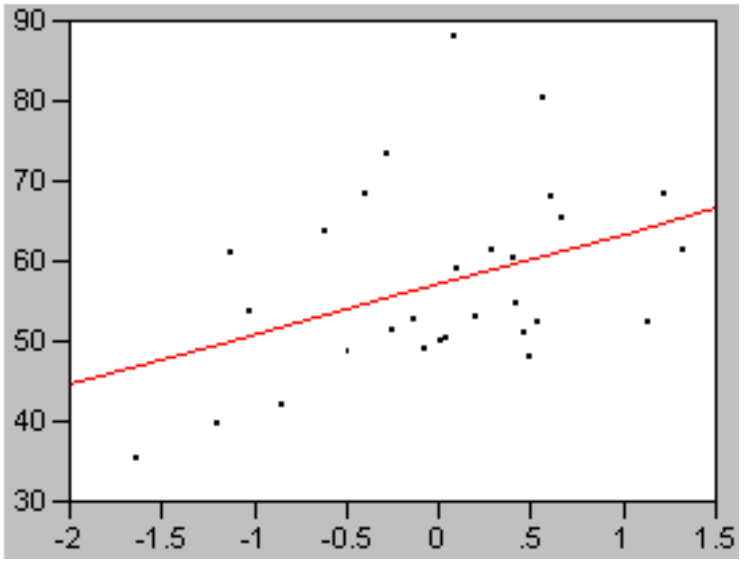
Linear regression for prediction

- slope and intercept define equation of line: **`Y = slope*X + intercept`**
- predict the value of **`Y`** for any value of **`X`**



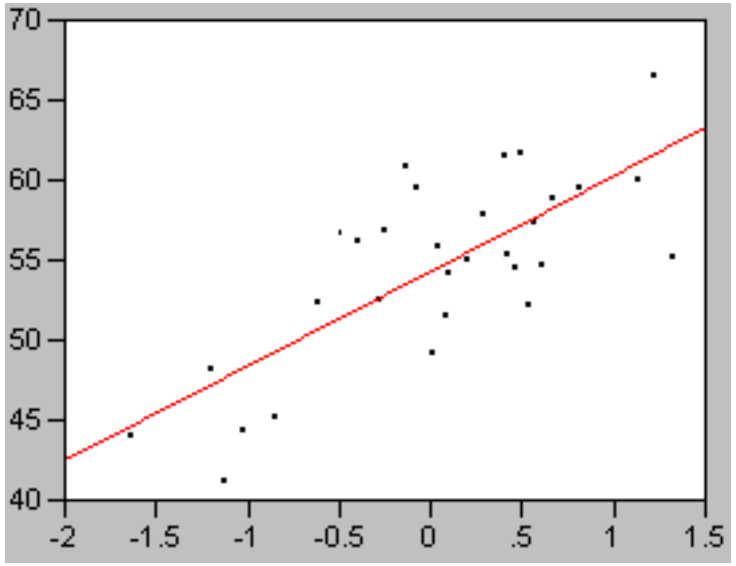
What is R^2 ?

- R^2 is a measure of how tightly the points fit around the best fit line
- The same best fit line can describe a variety of datasets **y= 6 x + 55**
- R2 tells us the fraction of the variance explained by the model.
- The predictive ability of a line depends on the R^2 value



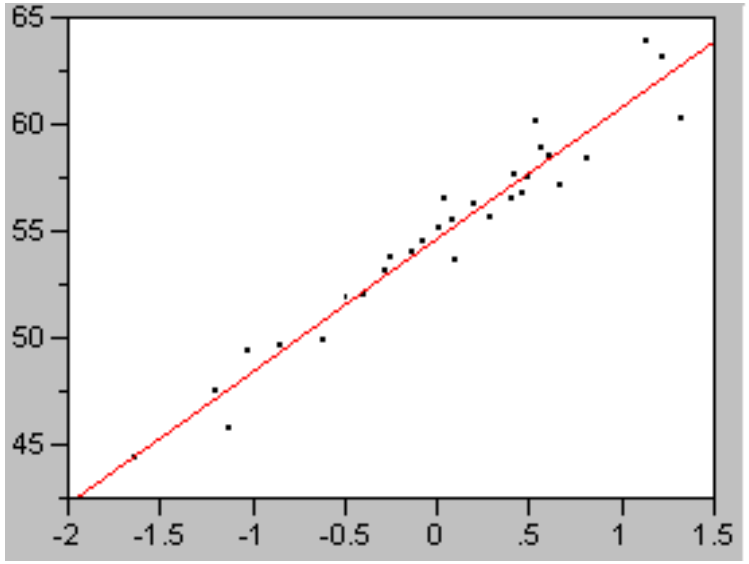
$R^2 = 0.16$

extremely poor predictive ability



$R^2 = 0.56$

moderate predictive ability



$R^2 = 0.94$

excellent predictive ability

Which kind of R^2?

- R reports "Multiple R-squared" and "Adjusted R-squared"
- For a simple linear regression, the numbers are usually similar.
- Which to report may depend on your field.
- When in doubt, report the adjusted R^2

Linear regression as a statistical test

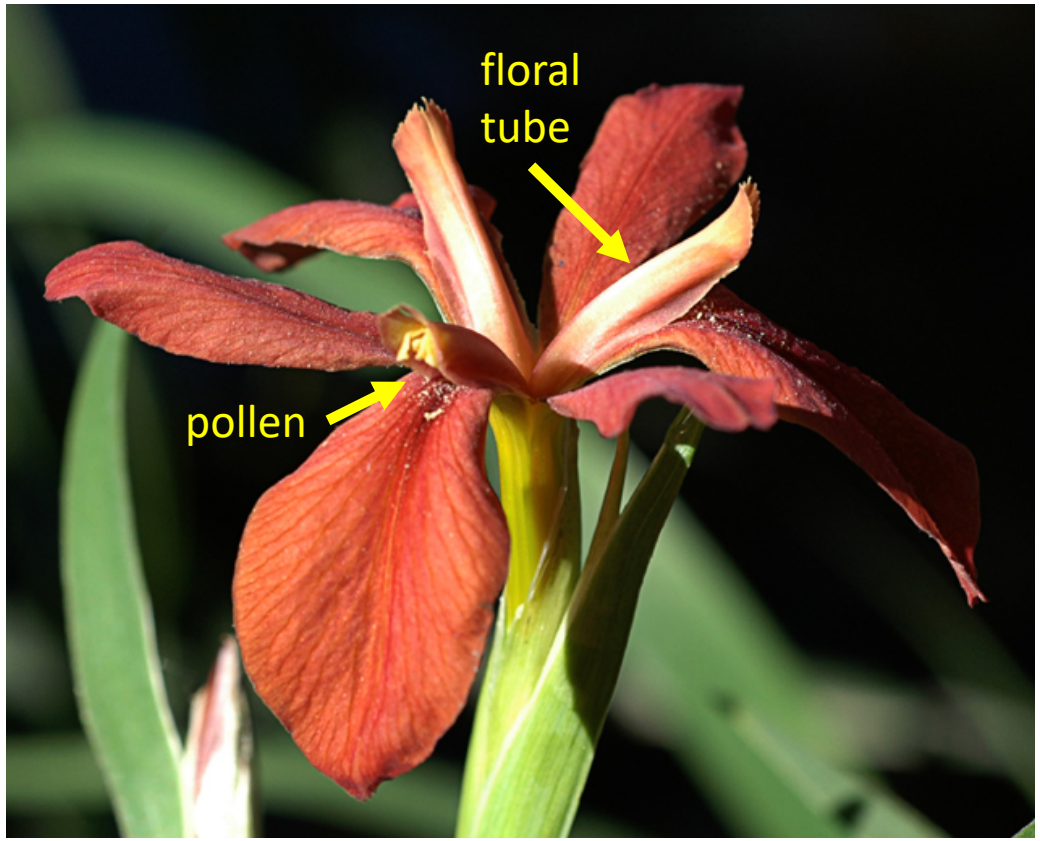
Jean & Alexander Heard
LIBRARIES

What does linear regression test?

- Linear regression tests whether the independent variable (X) has a significant effect on the dependent variable (Y)
- Mathematically: X has a significant effect on Y when the slope of the best fit line differs significantly from zero.
- Null hypothesis: the slope is zero.
- P assesses the probability that random variability in the data cause the slope to differ from zero.

Example: pollen vs. floral tube length

- independent variable: length of floral tube of an iris species
- dependent variable: pollen grains received
- Is a regression appropriate?
- Note: pollen grains received is counts.



Iris fulva photo by Ron Thomas © 2014 CC BY-NC-SA
http://bioimages.vanderbilt.edu/thomas/0031-00-03

Example data from Whitlock and Schluter (2nd ed.)
chapter 17. https://whitlockschluter.zoology.ubc.ca/r-code/rcode17

Assumptions of the linear regression test

- X and Y are independent (examine the design)
- The relationship is linear (vs. some other curve; examine the data)
- The **residuals** are normally distributed
- The variance of the **residuals** are the same for all values of X

What are residuals?

- Residuals are the distance from each point to the best-fit line.
- The least squares method minimizes the residuals.

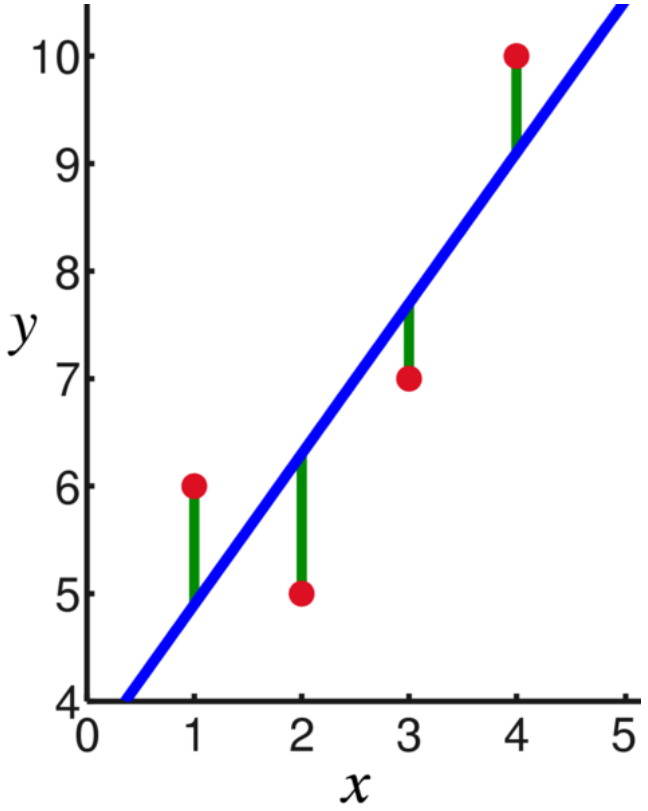
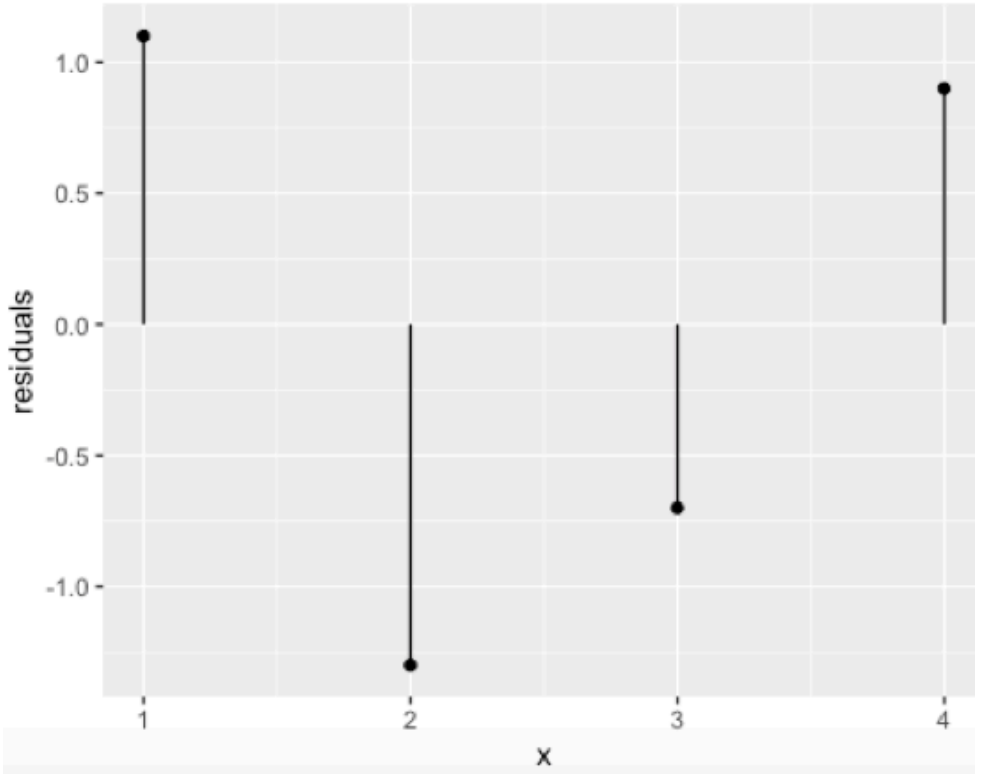


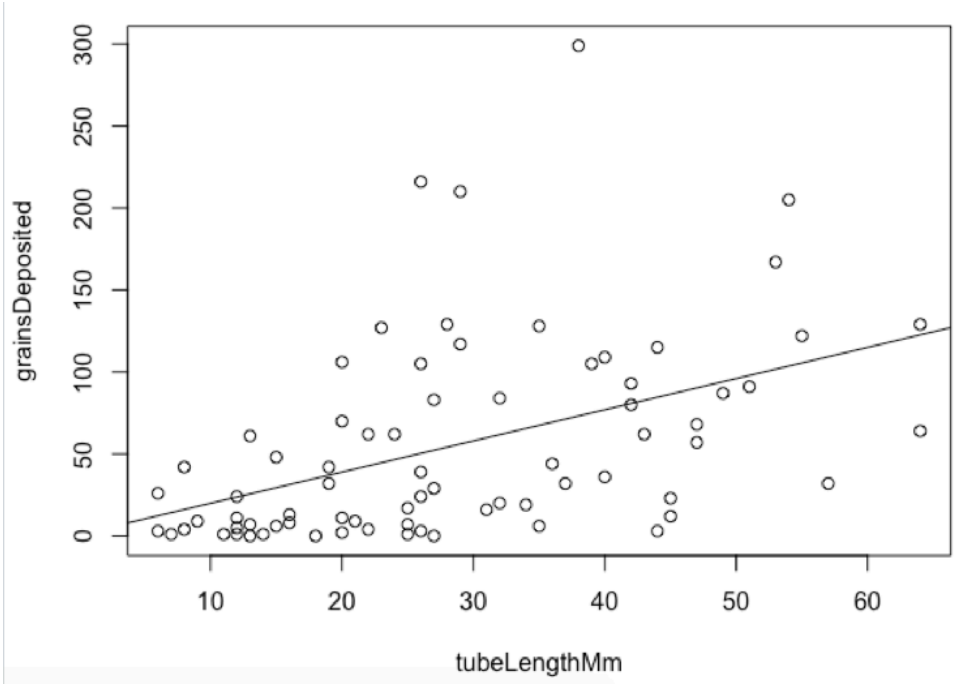
Image: Oleg Alexandrov Wikimedia Commons public domain

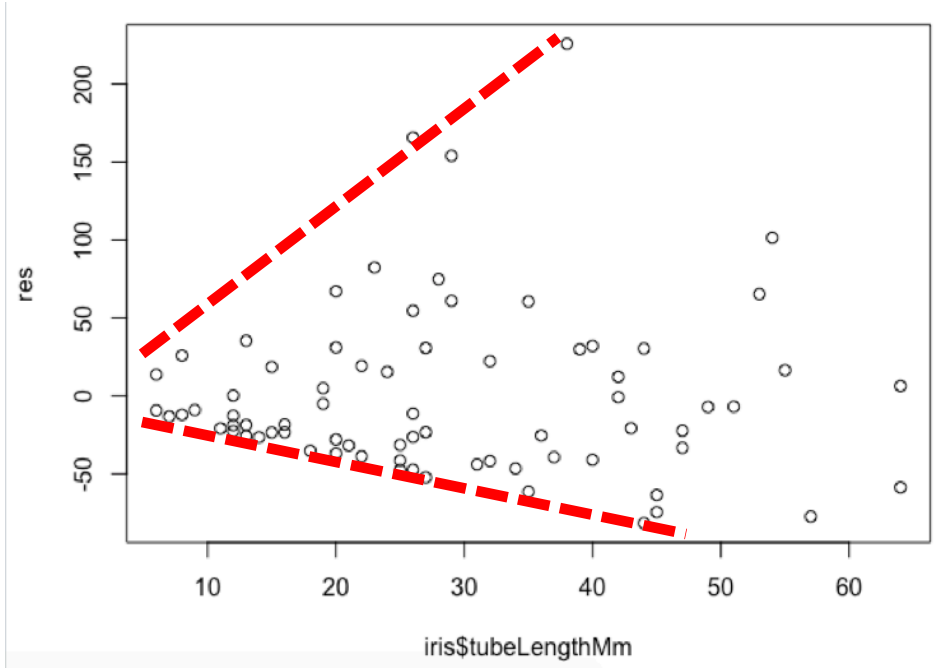


plot of residuals

Examine the data

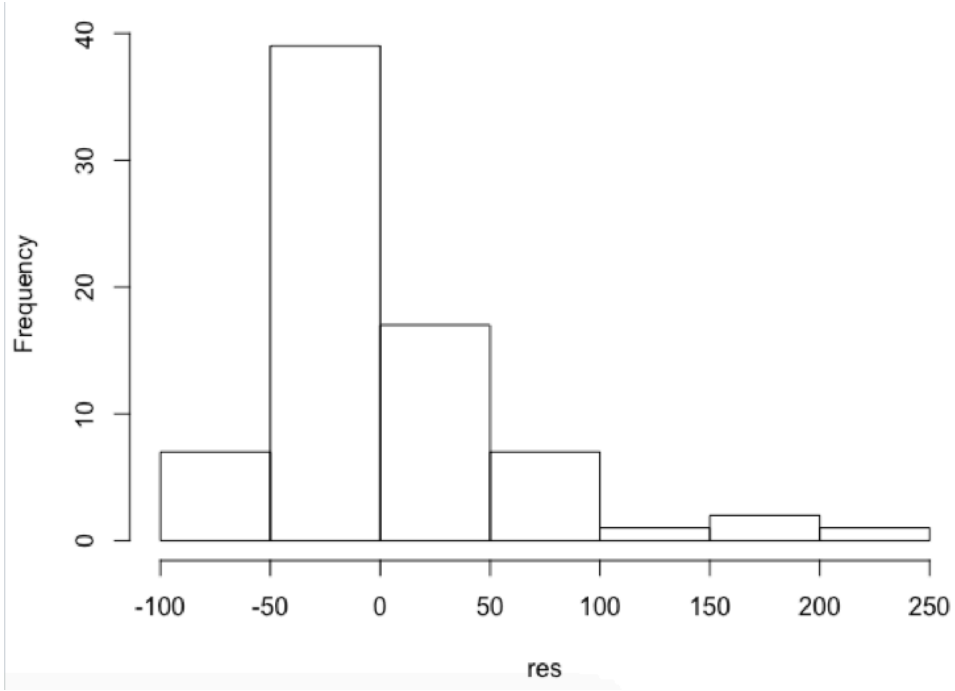
- Trend is increasing
- Residuals are problematic:
  - variance definitely NOT the same for all values of X
  - don't know about normality

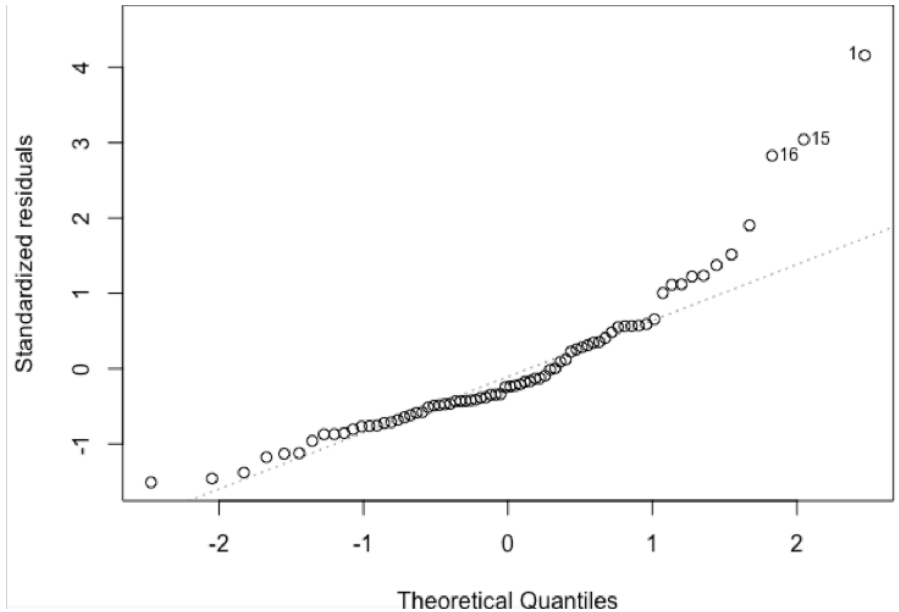




Residuals: untransformed

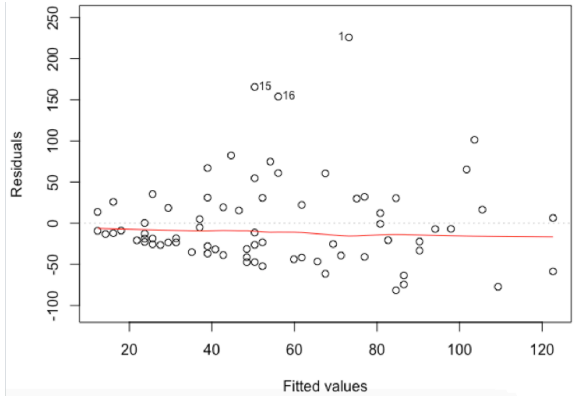
- Distribution skewed to right
- Use **`plot(linear_model)`** to generate normal quantile plot of residuals
- Shapiro-Wilkes test: P = 1.363e-06
- Y are counts; suggests square root transformation
- Must add a constant if negative numbers

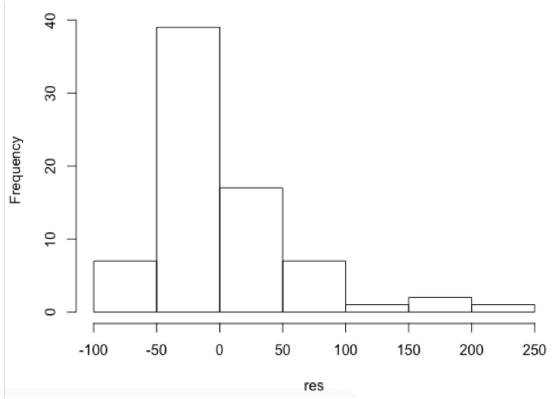


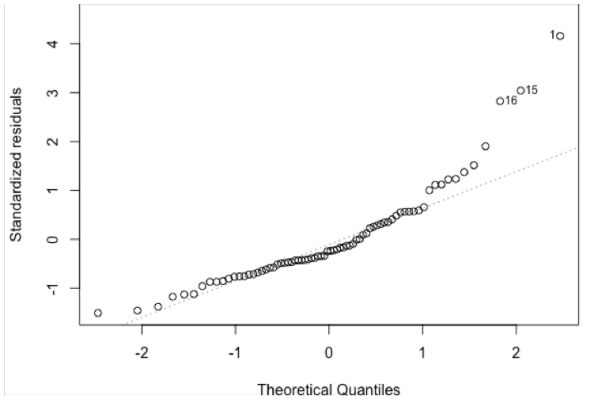


Before/after square root transformation

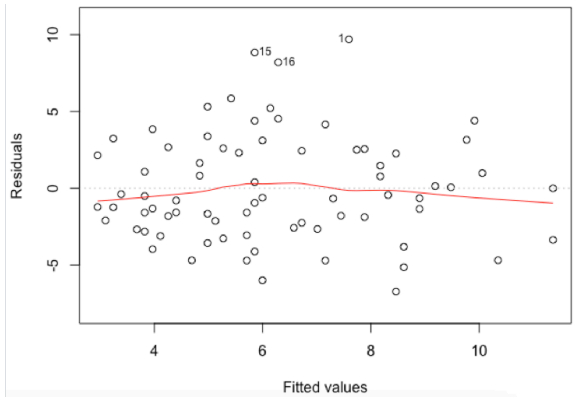
- Transformation greatly improves problems with assumptions
- Robust to deviations, but visible examination for outliers and non-homogeneous variance important https://doi.org/10.1101/498931

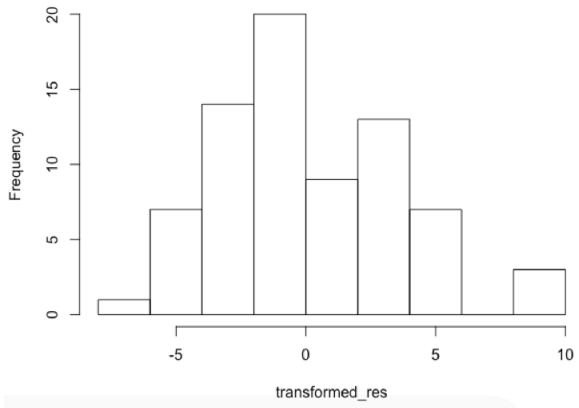


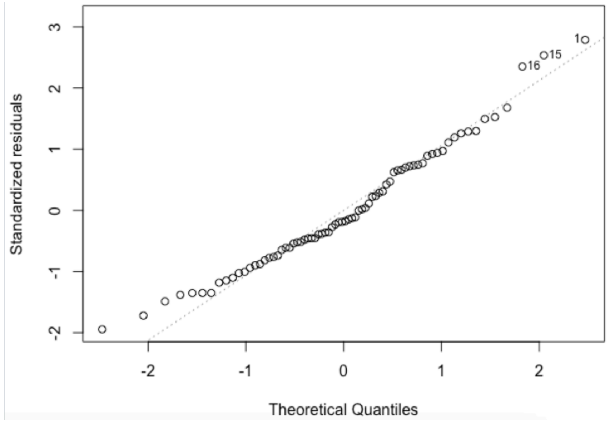




P = 1.363e-06







P = 0.1036

Shapiro-Wilkes test

Test after transformation

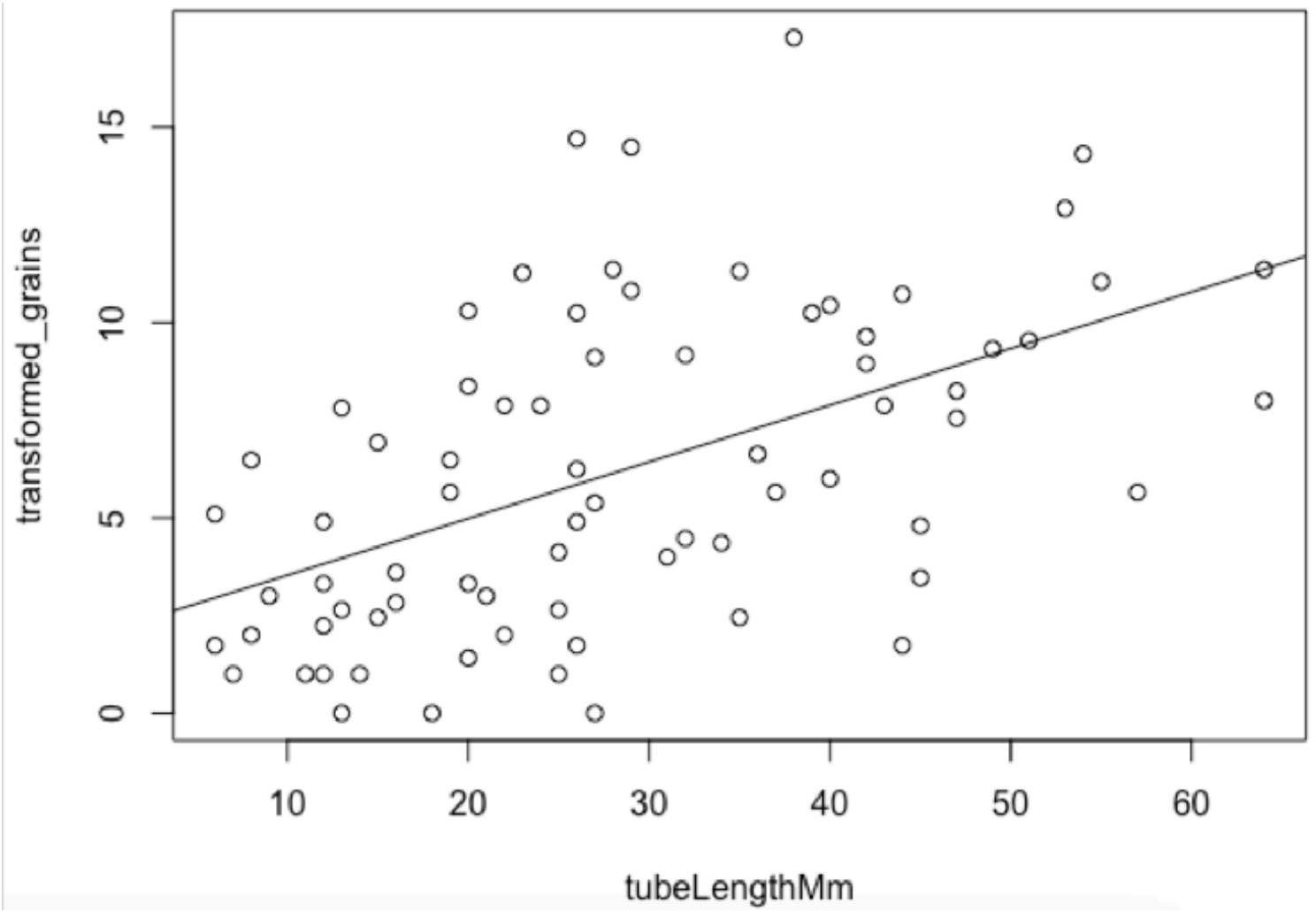
P = 1.9e-06

adjusted R^2 = 0.2617

Relationship is highly significant.

Not a tight fit to the line.

Large sample size allows detection of effect.



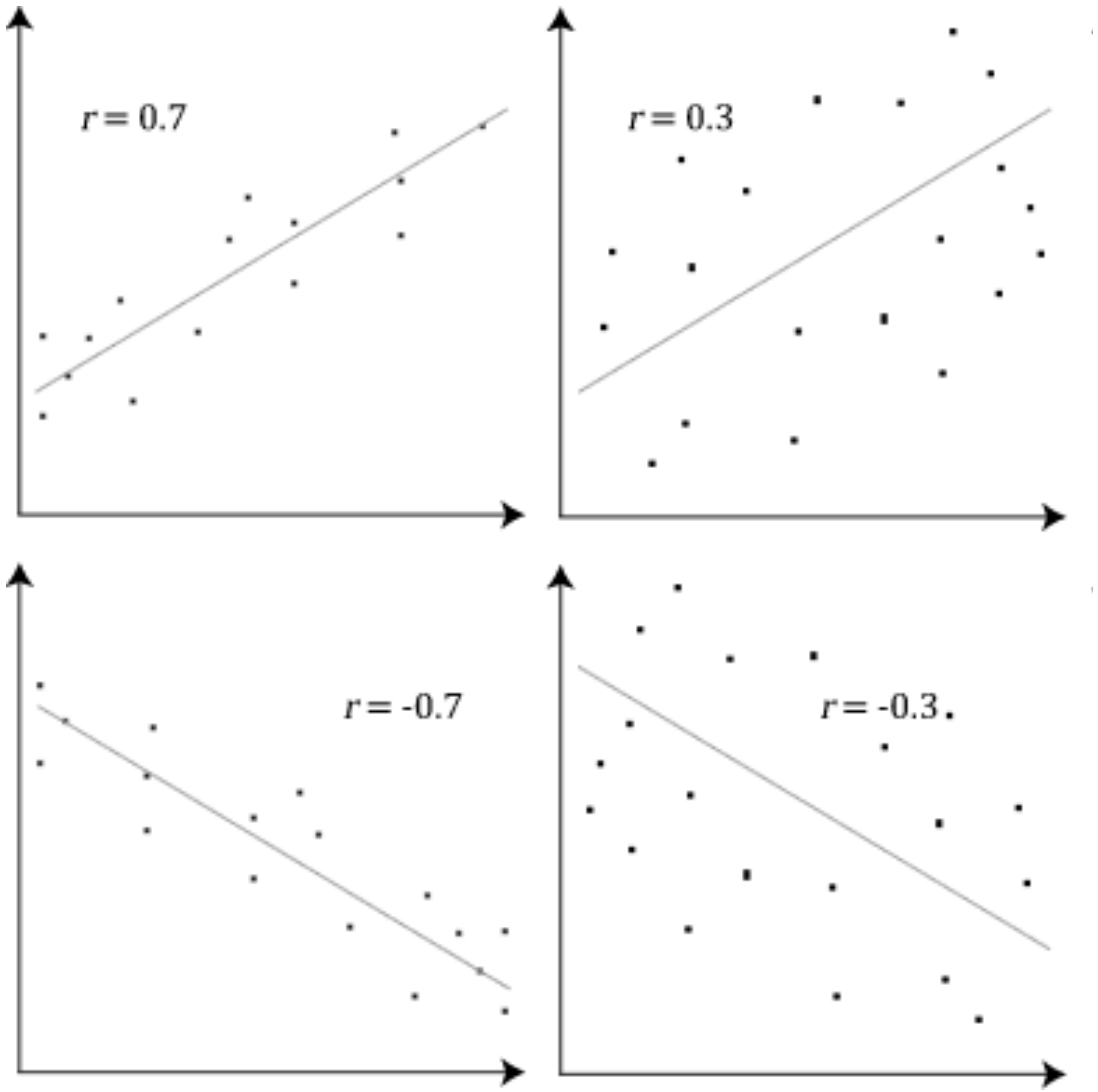
Correlation

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Correlation

- does not assume cause and effect
- assesses whether two continuous variables are related
- can be positive or negative
- assessed by correlation coefficient R



Assumptions of correlation

- Random sample
- Bivariate normal distribution
  - linear X/Y relationship
  - scatterplot is elliptical
  - X and Y distributions are separately normal
- The MVN library tests for multivariate normality

Non-parametric alternatives to correlation

- Kendall rank correlation test
  - more robust and less sensitive to error

  cor.test(v1, v2, method="kendall")
- Spearman rank correlation test
  - in common use

  cor.test(v1, v2, method="spearman")
- Usually the tests give similar results