

# REGRESSIONS

Correlations for the Advanced?



Erik Kusch

[erik.kusch@au.dk](mailto:erik.kusch@au.dk)

Section for Ecoinformatics & Biodiversity  
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)  
Aarhus University

27/01/2021

## 1 The Basics

- Correlation Tests
- Regression Models
- Least Squares vs. Maximum Likelihood

## 2 Methods & Models

- Single Linear Regression
- Mixed Effect Models
- Generalised Linear Models

## 3 Choosing the Right Method

## 1 The Basics

- Correlation Tests
- Regression Models
- Least Squares vs. Maximum Likelihood

## 2 Methods & Models

- Single Linear Regression
- Mixed Effect Models
- Generalised Linear Models

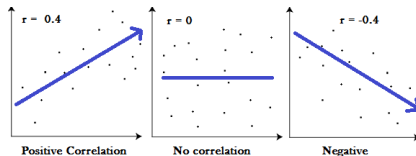
## 3 Choosing the Right Method

# Terminology

Correlation is **not** necessarily **causation** (spurious correlations).

Correlation tests yield two measurements:

- $r$  value (measure of correlation)
  - $r \approx 1$  (strong, positive correlation)
  - $r \approx 0$  (no correlation)
  - $r \approx -1$  (strong, negative correlation)
- $p$  value (measure of statistical significance)



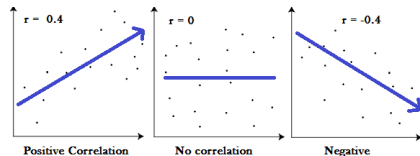
→ Get a feeling for it on [Guess The Correlation](#).

# Terminology

Correlation is **not** necessarily **causation** (spurious correlations).

Correlation tests yield two measurements:

- $r$  value (measure of correlation)
  - $r \approx 1$  (strong, positive correlation)
  - $r \approx 0$  (no correlation)
  - $r \approx -1$  (strong, negative correlation)
- $p$  value (measure of statistical significance)



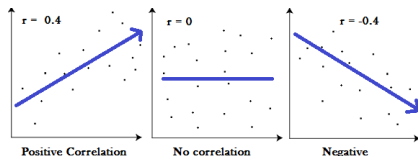
→ Get a feeling for it on [Guess The Correlation](#).

# Terminology

Correlation is **not** necessarily **causation** (spurious correlations).

Correlation tests yield two measurements:

- $r$  value (measure of correlation)
  - $r \approx 1$  (strong, positive correlation)
  - $r \approx 0$  (no correlation)
  - $r \approx -1$  (strong, negative correlation)
- $p$  value (measure of statistical significance)



→ Get a feeling for it on [Guess The Correlation](#).

# Types of Correlations

These approaches are extremely useful in data exploration and for preliminary analyses!

Prominent correlation tests include:

- Contingency Coefficient
- Kendall's Tau
- Spearman Correlation
- Pearson Correlation
- Cramer's V
- ANalysis Of VAriance (ANOVA)
- ...

**When you realize that all frequentist analyses are merely different versions of a correlation**



# Types of Correlations

These approaches are extremely useful in data exploration and for preliminary analyses!

Prominent correlation tests include:

- Contingency Coefficient
- Kendall's Tau
- Spearman Correlation
- Pearson Correlation
- Cramer's V
- ANalysis Of VAriance (ANOVA)
- ...

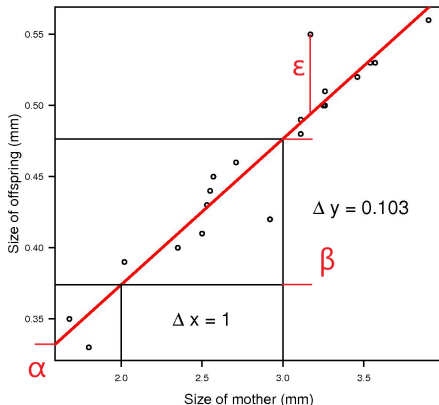
**When you realize that all frequentist analyses are merely different versions of a correlation**





# Terminology

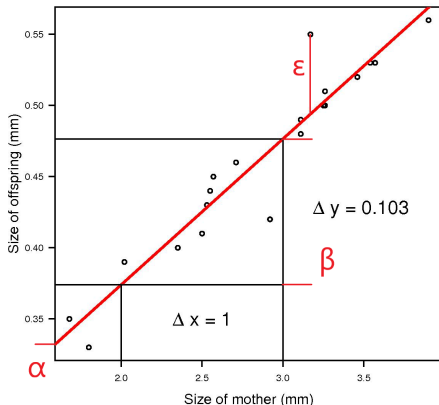
- $\alpha$  - The **Intercept**. The value of  $y$  when  $x = 0$  (also referred to as  $\beta_0$ ).
- $\beta_i$  - The **Correlation Coefficient**. The increase in  $y$  for a one-unit increase in dependent variable  $i$  (usually,  $x$  if only one dependent variable).
- $\epsilon$  - The **Random Error**. The deviation of data points from the regression line. Usually assumed to follow  $\epsilon \sim N(0, \sigma^2)$



Modified after Knut Helge Jensen.

# Terminology

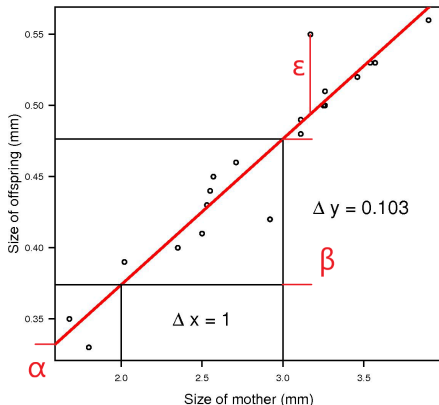
- $\alpha$  - The **Intercept**. The value of  $y$  when  $x = 0$  (also referred to as  $\beta_0$ ).
- $\beta_i$  - The **Correlation Coefficient**. The increase in  $y$  for a one-unit increase in dependent variable  $i$  (usually,  $x$  if only one dependent variable).
- $\epsilon$  - The **Random Error**. The deviation of data points from the regression line. Usually assumed to follow  $\epsilon \sim N(0, \sigma^2)$



Modified after Knut Helge Jensen.

# Terminology

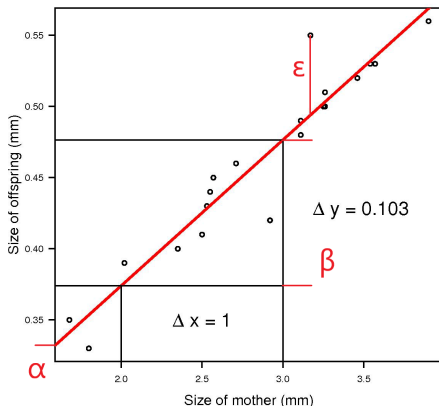
- $\alpha$  - The **Intercept**. The value of  $y$  when  $x = 0$  (also referred to as  $\beta_0$ ).
- $\beta_i$  - The **Correlation Coefficient**. The increase in  $y$  for a one-unit increase in dependent variable  $i$  (usually,  $x$  if only one dependent variable).
- $\epsilon$  - The **Random Error**. The deviation of data points from the regression line. Usually assumed to follow  $\epsilon \sim N(0, \sigma^2)$



Modified after Knut Helge Jensen.

# Terminology

- $\alpha$  - The **Intercept**. The value of  $y$  when  $x = 0$  (also referred to as  $\beta_0$ ).
- $\beta_i$  - The **Correlation Coefficient**. The increase in  $y$  for a one-unit increase in dependent variable  $i$  (usually,  $x$  if only one dependent variable).
- $\epsilon$  - The **Random Error**. The deviation of data points from the regression line. Usually assumed to follow  $\epsilon \sim N(0, \sigma^2)$



Modified after Knut Helge Jensen.

# Assumptions in Theory

Linear regression models need to be inspected for violations of assumptions after regressing:

- **Linearity** - *Residuals vs. Fitted values*

Non-linear patterns identify a non-linear relationship between dependent and independent variables.

- **Normality** - *Normal Q-Q plot*

Non-normal distribution of residuals shows that the assumption of  $\epsilon \sim N(0, \sigma^2)$  is violated.

- **Homoscedasticity** - *Scale Location*

Non-constant variance shows that the assumption of homoscedasticity (invoked by least squares fitting) is violated.

- **Outliers** - *Residuals vs. Leverage*

A non-zero trend identifies the presence of influential outliers.

# Assumptions in Theory

Linear regression models need to be inspected for violations of assumptions after regressing:

- **Linearity** - *Residuals vs. Fitted values*

Non-linear patterns identify a non-linear relationship between dependent and independent variables.

- **Normality** - *Normal Q-Q plot*

Non-normal distribution of residuals shows that the assumption of  $\epsilon \sim N(0, \sigma^2)$  is violated.

- **Homoscedasticity** - *Scale Location*

Non-constant variance shows that the assumption of homoscedasticity (invoked by least squares fitting) is violated.

- **Outliers** - *Residuals vs. Leverage*

A non-zero trend identifies the presence of influential outliers.

# Assumptions in Theory

Linear regression models need to be inspected for violations of assumptions after regressing:

- **Linearity** - *Residuals vs. Fitted values*

Non-linear patterns identify a non-linear relationship between dependent and independent variables.

- **Normality** - *Normal Q-Q plot*

Non-normal distribution of residuals shows that the assumption of  $\epsilon \sim N(0, \sigma^2)$  is violated.

- **Homoscedasticity** - *Scale Location*

Non-constant variance shows that the assumption of homoscedasticity (invoked by least squares fitting) is violated.

- **Outliers** - *Residuals vs. Leverage*

A non-zero trend identifies the presence of influential outliers.

# Assumptions in Theory

Linear regression models need to be inspected for violations of assumptions after regressing:

- **Linearity** - *Residuals vs. Fitted values*

Non-linear patterns identify a non-linear relationship between dependent and independent variables.

- **Normality** - *Normal Q-Q plot*

Non-normal distribution of residuals shows that the assumption of  $\epsilon \sim N(0, \sigma^2)$  is violated.

- **Homoscedasticity** - *Scale Location*

Non-constant variance shows that the assumption of homoscedasticity (invoked by least squares fitting) is violated.

- **Outliers** - *Residuals vs. Leverage*

A non-zero trend identifies the presence of influential outliers.



# Assumptions in Theory

Linear regression models need to be inspected for violations of assumptions after regressing:

- **Linearity** - *Residuals vs. Fitted values*

Non-linear patterns identify a non-linear relationship between dependent and independent variables.

- **Normality** - *Normal Q-Q plot*

Non-normal distribution of residuals shows that the assumption of  $\epsilon \sim N(0, \sigma^2)$  is violated.

- **Homoscedasticity** - *Scale Location*

Non-constant variance shows that the assumption of homoscedasticity (invoked by least squares fitting) is violated.

- **Outliers** - *Residuals vs. Leverage*

A non-zero trend identifies the presence of influential outliers.

# Linearity

**"Linear" does not mean that there is a straight line!**

"Linear" models refer to:

- *Additive* effects of predictors ( $P$ )

$$\beta_1 P_1 + \beta_2 P_2 + \dots + \beta_k P_k = \sum_i^k \beta_i P_i$$

- *Deterministic* combination of predictors ( $P$ ) to defining response ( $R$ )

$$R = \alpha + \epsilon + \sum_i^k \beta_i P_i$$

You can model curved relationships with a number of *tricks* like **exponential functions, polynomials, splines**.

# Linearity

**"Linear" does not mean that there is a straight line!**

"Linear" models refer to:

- *Additive* effects of predictors ( $P$ )

$$\beta_1 P_1 + \beta_2 P_2 + \dots + \beta_k P_k = \sum_i^k \beta_i P_i$$

- *Deterministic* combination of predictors ( $P$ ) to defining response ( $R$ )

$$R = \alpha + \epsilon + \sum_i^k \beta_i P_i$$

You can model curved relationships with a number of *tricks* like **exponential functions, polynomials, splines**.

# Linearity

**"Linear" does not mean that there is a straight line!**

"Linear" models refer to:

- *Additive* effects of predictors ( $P$ )

$$\beta_1 P_1 + \beta_2 P_2 + \dots + \beta_k P_k = \sum_i^k \beta_i P_i$$

- *Deterministic* combination of predictors ( $P$ ) to defining response ( $R$ )

$$R = \alpha + \epsilon + \sum_i^k \beta_i P_i$$

You can model curved relationships with a number of *tricks* like **exponential functions, polynomials, splines**.

# Linearity

**"Linear" does not mean that there is a straight line!**

"Linear" models refer to:

- *Additive* effects of predictors ( $P$ )

$$\beta_1 P_1 + \beta_2 P_2 + \dots + \beta_k P_k = \sum_i^k \beta_i P_i$$

- *Deterministic* combination of predictors ( $P$ ) to defining response ( $R$ )

$$R = \alpha + \epsilon + \sum_i^k \beta_i P_i$$

You can model curved relationships with a number of *tricks* like **exponential functions, polynomials, splines**.

# Linearity

**"Linear" does not mean that there is a straight line!**

"Linear" models refer to:

- *Additive* effects of predictors ( $P$ )

$$\beta_1 P_1 + \beta_2 P_2 + \dots + \beta_k P_k = \sum_i^k \beta_i P_i$$

- *Deterministic* combination of predictors ( $P$ ) to defining response ( $R$ )

$$R = \alpha + \epsilon + \sum_i^k \beta_i P_i$$

You can model curved relationships with a number of *tricks* like **exponential functions, polynomials, splines**.

# Normality

## Normal distributions are common in nature.

Normal/Gaussian distributions are important because:

- Easy to do Math with
- Common in nature:
  - Processes which add fluctuations create dampening
  - Dampened fluctuations create Gaussian distributions
- Information reduced to mean and variance
- Underlying process can't be inferred from the distribution

Models which are **mechanisitically wrong** can still excell in **predictive power**.

# Normality

**Normal distributions are common in nature.**

Normal/Gaussian distributions are important because:

- Easy to do Math with
- Common in nature:
  - Processes which add fluctuations create dampening
  - Dampened fluctuations create Gaussian distributions
- Information reduced to mean and variance
- Underlying process can't be inferred from the distribution

Models which are **mechanisitically wrong** can still excell in **predictive power**.



# Normality

**Normal distributions are common in nature.**

Normal/Gaussian distributions are important because:

- Easy to do Math with
- Common in nature:
  - Processes which add fluctuations create dampening
  - Dampened fluctuations create Gaussian distributions
- Information reduced to mean and variance
- Underlying process can't be inferred from the distribution

Models which are **mechanisitically wrong** can still excell in **predictive power**.

# Normality

**Normal distributions are common in nature.**

Normal/Gaussian distributions are important because:

- Easy to do Math with
- Common in nature:
  - Processes which add fluctuations create dampening
  - Dampened fluctuations create Gaussian distributions
- Information reduced to mean and variance
- Underlying process can't be inferred from the distribution

Models which are **mechanisitically wrong** can still excell in **predictive power**.

# Normality

**Normal distributions are common in nature.**

Normal/Gaussian distributions are important because:

- Easy to do Math with
- Common in nature:
  - Processes which add fluctuations create dampening
  - Dampened fluctuations create Gaussian distributions
- Information reduced to mean and variance
- Underlying process can't be inferred from the distribution

Models which are **mechanisitically wrong** can still excell in **predictive power**.

# Normality

**Normal distributions are common in nature.**

Normal/Gaussian distributions are important because:

- Easy to do Math with
- Common in nature:
  - Processes which add fluctuations create dampening
  - Dampened fluctuations create Gaussian distributions
- Information reduced to mean and variance
- Underlying process can't be inferred from the distribution

Models which are **mechanisitically wrong** can still excell in **predictive power**.

# Normality

**Normal distributions are common in nature.**

Normal/Gaussian distributions are important because:

- Easy to do Math with
- Common in nature:
  - Processes which add fluctuations create dampening
  - Dampened fluctuations create Gaussian distributions
- Information reduced to mean and variance
- Underlying process can't be inferred from the distribution

Models which are **mechanisitically wrong** can still excell in **predictive power**.

# Homoscedasticity

**Variance contains important information.**

Variation is at the heart of biology:

- *Abiotic* variation governs biological processes and patterns
- *Biological* variation can alter how ecosystem function and persist

Ignoring variance is a **cardinal sin**.

# Homoscedasticity

**Variance contains important information.**

Variation is at the heart of biology:

- *Abiotic* variation governs biological processes and patterns
- *Biological* variation can alter how ecosystem function and persist

Ignoring variance is a **cardinal sin**.

# Homoscedasticity

**Variance contains important information.**

Variation is at the heart of biology:

- *Abiotic* variation governs biological processes and patterns
- *Biological* variation can alter how ecosystem function and persist

Ignoring variance is a **cardinal sin**.



# Homoscedasticity

**Variance contains important information.**

Variation is at the heart of biology:

- *Abiotic* variation governs biological processes and patterns
- *Biological* variation can alter how ecosystem function and persist

Ignoring variance is a **cardinal sin**.

# Homoscedasticity

**Variance contains important information.**

Variation is at the heart of biology:

- *Abiotic* variation governs biological processes and patterns
- *Biological* variation can alter how ecosystem function and persist

Ignoring variance is a **cardinal sin**.

# Outliers

**Unexpected data values can still be valuable.**

■ *Outlier sources:*

- Decimal delimiter errors
- Imprecise recording
- Noise
- Extreme events/occurrences/records
- Surprising findings

■ *Identify outliers:*

- *Studentised residuals* ( $|\frac{e_i}{s d_e}|$ )
- *Knowledge and Intuition* of the biological study system and sampling units

# Outliers

**Unexpected data values can still be valuable.**

■ *Outlier sources:*

- Decimal delimiter errors
- Imprecise recording
- Noise
- Extreme events/occurrences/records
- Surprising findings

■ *Identify outliers:*

- *Studentised residuals* ( $|\frac{e_i}{s d_e}|$ )
- *Knowledge and Intuition* of the biological study system and sampling units

# Outliers

**Unexpected data values can still be valuable.**

■ *Outlier sources:*

- Decimal delimiter errors
- Imprecise recording
- Noise
- Extreme events/occurrences/records
- Surprising findings

■ *Identify outliers:*

- *Studentised residuals* ( $|\frac{e_i}{s d_e}|$ )
- *Knowledge and Intuition* of the biological study system and sampling units

# Outliers

**Unexpected data values can still be valuable.**

■ *Outlier sources:*

- Decimal delimiter errors
- Imprecise recording
- Noise
- Extreme events/occurrences/records
- Surprising findings

■ *Identify outliers:*

- *Studentised residuals* ( $|\frac{e_i}{sd_e}|$ )
- *Knowledge and Intuition* of the biological study system and sampling units

# Outliers

**Unexpected data values can still be valuable.**

■ *Outlier sources:*

- Decimal delimiter errors
- Imprecise recording
- Noise
- Extreme events/occurrences/records
- Surprising findings

■ *Identify outliers:*

- *Studentised residuals* ( $|\frac{e_i}{sd_e}|$ )
- *Knowledge and Intuition* of the biological study system and sampling units

# Outliers

**Unexpected data values can still be valuable.**

■ *Outlier sources:*

- Decimal delimiter errors
- Imprecise recording
- Noise
- Extreme events/occurrences/records
- Surprising findings

■ *Identify outliers:*

- *Studentised residuals* ( $|\frac{e_i}{sde}|$ )
- *Knowledge and Intuition* of the biological study system and sampling units



# Outliers

**Unexpected data values can still be valuable.**

■ *Outlier sources:*

- Decimal delimiter errors
- Imprecise recording
- Noise
- Extreme events/occurrences/records
- Surprising findings

■ *Identify outliers:*

- *Studentised residuals* ( $|\frac{e_i}{s d_e}|$ )
- *Knowledge and Intuition* of the biological study system and sampling units

# Assumptions in R

Simply type `plot(...)` with `...` denoting your regression model.

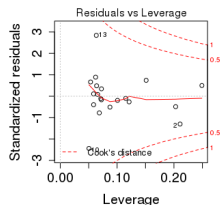
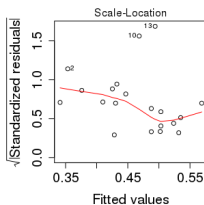
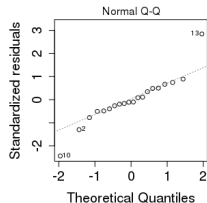
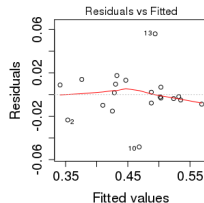
You can also target individual plots:

1. `plot(..., 1)` (Residuals/Fitted)

2. `plot(..., 2)` (Normal Q-Q plot)

3. `plot(..., 3)` (Scale Location)

4. `plot(..., 4)`  
(Residuals/Leverage)



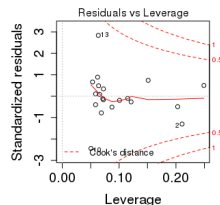
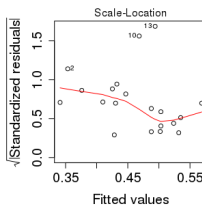
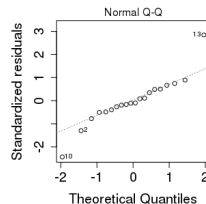
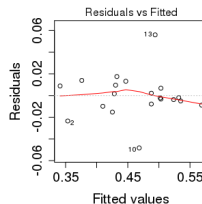
By Knut Helge Jensen.

# Assumptions in R

Simply type `plot(...)` with `...` denoting your regression model.

You can also target individual plots:

1. `plot(..., 1)` (Residuals/Fitted)
2. `plot(..., 2)` (Normal Q-Q plot)
3. `plot(..., 3)` (Scale Location)
4. `plot(..., 4)`  
(Residuals/Leverage)



By Knut Helge Jensen.

# Types of Regressions

Less model variables result in a more interpretable model!

Prominent regression approaches include the following:

- **Single Linear Regression**
- Multiple Linear Regression
- **Linear Mixed Effect Models**
- **Generalized Linear Models**

- Polynomial Regressions
- Generalized Additive Models
- Regression Splines
- Smoothing Splines
- Local Regressions
- ...

# Types of Regressions

Less model variables result in a more interpretable model!

Prominent regression approaches include the following:

- **Single Linear Regression**
- Multiple Linear Regression
- **Linear Mixed Effect Models**
- **Generalized Linear Models**

- Polynomial Regressions
- Generalized Additive Models
- Regression Splines
- Smoothing Splines
- Local Regressions
- ...

# Types of Regressions

Less model variables result in a more interpretable model!

Prominent regression approaches include the following:

- **Single Linear Regression**
- Multiple Linear Regression
- **Linear Mixed Effect Models**
- **Generalized Linear Models**

- Polynomial Regressions
- Generalized Additive Models
- Regression Splines
- Smoothing Splines
- Local Regressions
- ...

# Least Squares vs. Maximum Likelihood

These methods refer to **parameter estimation**.

*Ordinary Least Squares (OSL):*

- Used for most basic linear regressions
- obtain coefficient estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  such that the linear model fits the available data well — that is, so that  $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$  for  $i = 1, \dots, n$ .

Minimize:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

with  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

*Maximum Likelihood Estimation (MLE):*

- Used in logistic regressions and generalized linear models
- estimates for  $\beta_0$  and  $\beta_1$  such that the predicted probability  $\hat{Pr}(x_j)$  corresponds to the observed response variable status.

Maximize:

$$\ell(\theta) = \prod_{i=1}^n f(x_i | \theta) \quad (2)$$

## 1 The Basics

- Correlation Tests
- Regression Models
- Least Squares vs. Maximum Likelihood

## 2 Methods & Models

- Single Linear Regression
- Mixed Effect Models
- Generalised Linear Models

## 3 Choosing the Right Method



# Purpose & Assumptions

## Single linear regression

`lm()` in base R

*Purpose:* Identify whether and how two variables are related.

■ Down to *Study-Design*:

- Response variable is continuous (ratio or interval scale)
- Predictor variable is continuous (ratio or interval scale)
- Variable values are **independent** (not paired)

*Assumptions:*

■ Need for *Post-Hoc Tests*:

- Variable values follow **homoscedasticity** (equal variance across entire data range)
- Residuals follow normal distribution (**normality**)
- Absence of **influential outliers**
- Response and Predictor are related in a **linear** fashion

# Purpose & Assumptions

## Single linear regression

`lm()` in base R

*Purpose:* Identify whether and how two variables are related.

■ Down to *Study-Design*:

- Response variable is continuous (ratio or interval scale)
- Predictor variable is continuous (ratio or interval scale)
- Variable values are **independent** (not paired)

*Assumptions:*

■ Need for *Post-Hoc Tests*:

- Variable values follow **homoscedasticity** (equal variance across entire data range)
- Residuals follow normal distribution (**normality**)
- Absence of **influential outliers**
- Response and Predictor are related in a **linear** fashion

# Example - The Data

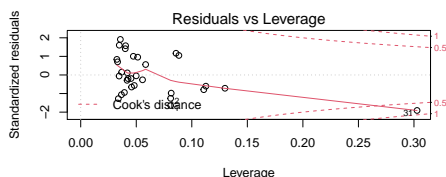
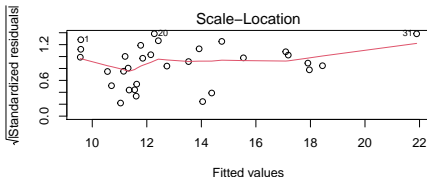
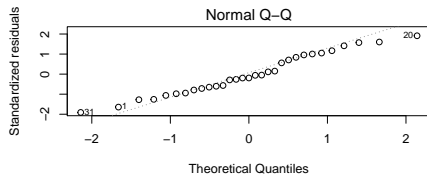
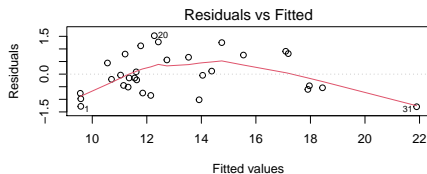
```
# measures of Diameter (labelled as Girth), Height, and Volume of Timber  
data("trees")  
head(trees)
```

##	Girth	Height	Volume
## 1	8.3	70	10.3
## 2	8.6	65	10.3
## 3	8.8	63	10.2
## 4	10.5	72	16.4
## 5	10.7	81	18.8
## 6	10.8	83	19.7

→ Let's see if there is a good regression to be had between *Girth* and *Volume*.

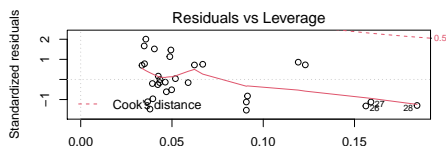
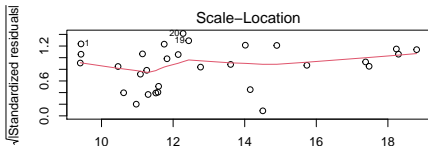
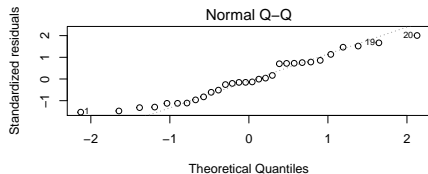
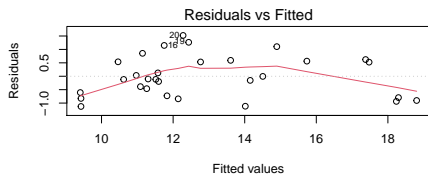
# Example - The Model

```
SingleLin_Mod <- with(trees, lm(Girth ~ Volume))
par(mfrow=c(2,2))
plot(SingleLin_Mod)
```



# Example - Refining The Model

```
trees <- trees[-31,] # removing the influential outlier in row 31
SingleLin_Mod <- with(trees, lm(Girth ~ Volume))
par(mfrow=c(2,2))
plot(SingleLin_Mod)
```



# Example - Model Output

```
summary(SingleLin_Mod)
```

```
##
## Call:
## lm(formula = Girth ~ Volume)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.126 -0.699 -0.109  0.557  1.521
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.4141     0.3217    23.0  <2e-16 ***
## Volume        0.1954     0.0101    19.3  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.772 on 28 degrees of freedom
## Multiple R-squared:  0.93,    Adjusted R-squared:  0.928
## F-statistic: 374 on 1 and 28 DF,  p-value: <2e-16
```

At a Volume of 0, Girth is predicted to be 7.4141 (of course that doesn't make sense, not only is a volume of 0 biological nonsense, height also plays a part here for sure). For a one-unit increase in Volume, Girth is predicted to go up by 0.1954 inches (yes, they recorded in inches). Both estimates are statistically significant.

# Purpose & Assumptions

## Linear mixed effect model

`lme()` in base `nlme` package

*Purpose:* Identify whether and how variables are related.

- Down to *Study-Design*:

- Response variable is continuous (ratio or interval scale)
- Predictor variables are continuous (ratio or interval scale) and/or categorical (metric or ordinal scale)

*Assumptions:*

- Need for *Post-Hoc Tests*:

- Variable values follow **homoscedasticity** (equal variance across entire data range)
- Residuals follow normal distribution (**normality**)
- Absence of **influential outliers**
- Response and Predictor are related in a **linear** fashion

# Purpose & Assumptions

## Linear mixed effect model

`lme()` in base `nlme` package

*Purpose:* Identify whether and how variables are related.

■ Down to *Study-Design*:

- Response variable is continuous (ratio or interval scale)
- Predictor variables are continuous (ratio or interval scale) and/or categorical (metric or ordinal scale)

*Assumptions:*

■ Need for *Post-Hoc Tests*:

- Variable values follow **homoscedasticity** (equal variance across entire data range)
- Residuals follow normal distribution (**normality**)
- Absence of **influential outliers**
- Response and Predictor are related in a **linear** fashion



# Fixed vs. Random Effects

Fixed effects and random effects are also referred to as fixed effect factors and random effect factors.

## Fixed Effects:

- Informative factor levels regarding hypothesis.
- Want to study these levels and their effects.
- Factor levels are deliberate part of the study-design.
- Higher sample size  $\neq$  higher number of levels.

## Random Effects:

- Uninformative factor levels regarding hypothesis.
- Do not want to study these levels and their effects.
- Factor levels are imposed by nature/type of study.
- Usually: higher sample size = higher number of levels.

Usually stored in R in `factor` mode/class.

# Fixed vs. Random Effects

Fixed effects and random effects are also referred to as fixed effect factors and random effect factors.

## Fixed Effects:

- Informative factor levels regarding hypothesis.
- Want to study these levels and their effects.
- Factor levels are deliberate part of the study-design.
- Higher sample size  $\neq$  higher number of levels.

## Random Effects:

- Uninformative factor levels regarding hypothesis.
- Do not want to study these levels and their effects.
- Factor levels are imposed by nature/type of study.
- Usually: higher sample size = higher number of levels.

Usually stored in R in `factor` mode/class.

# Fixed vs. Random Effects

Fixed effects and random effects are also referred to as fixed effect factors and random effect factors.

## Fixed Effects:

- Informative factor levels regarding hypothesis.
- Want to study these levels and their effects.
- Factor levels are deliberate part of the study-design.
- Higher sample size  $\neq$  higher number of levels.

## Random Effects:

- Uninformative factor levels regarding hypothesis.
- Do not want to study these levels and their effects.
- Factor levels are imposed by nature/type of study.
- Usually: higher sample size = higher number of levels.

Usually stored in R in `factor` mode/class.

# Fixed vs. Random Effects

Fixed effects and random effects are also referred to as fixed effect factors and random effect factors.

## Fixed Effects:

- Informative factor levels regarding hypothesis.
- Want to study these levels and their effects.
- Factor levels are deliberate part of the study-design.
- Higher sample size  $\neq$  higher number of levels.

## Random Effects:

- Uninformative factor levels regarding hypothesis.
- Do not want to study these levels and their effects.
- Factor levels are imposed by nature/type of study.
- Usually: higher sample size = higher number of levels.

Usually stored in R in `factor` mode/class.

# Fixed vs. Random Effects

Fixed effects and random effects are also referred to as fixed effect factors and random effect factors.

## Fixed Effects:

- Informative factor levels regarding hypothesis.
- Want to study these levels and their effects.
- Factor levels are deliberate part of the study-design.
- Higher sample size  $\neq$  higher number of levels.

## Random Effects:

- Uninformative factor levels regarding hypothesis.
- Do not want to study these levels and their effects.
- Factor levels are imposed by nature/type of study.
- Usually: higher sample size = higher number of levels.

Usually stored in R in `factor` mode/class.

# Fixed vs. Random Effects

Fixed effects and random effects are also referred to as fixed effect factors and random effect factors.

## Fixed Effects:

- Informative factor levels regarding hypothesis.
- Want to study these levels and their effects.
- Factor levels are deliberate part of the study-design.
- Higher sample size  $\neq$  higher number of levels.

## Random Effects:

- Uninformative factor levels regarding hypothesis.
- Do not want to study these levels and their effects.
- Factor levels are imposed by nature/type of study.
- Usually: higher sample size = higher number of levels.

Usually stored in R in `factor` mode/class.

# Example - The Data

```
# measures of Weight, Diet, Time, and Chicks  
data("ChickWeight")  
head(ChickWeight)
```

```
##      weight  Time  Chick  Diet  
## 1         42     0      1     1  
## 2         51     2      1     1  
## 3         59     4      1     1  
## 4         64     6      1     1  
## 5         76     8      1     1  
## 6         93    10      1     1
```

→ Let's see if there is a good regression to be had between *weight* and *Time* while accounting for random effects belonging to *Chick*, and fixed effects of *Diet*.

# Example - The Model

```
library(nlme)
MultiLin_Base <- lme(weight ~ Time*Diet, # weight as an interaction of time and diet
  random = ~+1|Chick, # random effect of Chick
  data = ChickWeight)
```

We now have our model. However, we know that time is a component and we likely have repeated samples here. In these cases, we need to account for auto-correlation by defining a correlation structure.

```
MultiLin_Mod <- lme(weight ~ Time*Diet, random = ~+1|Chick,
  cor=corAR1(), # adding autocorrelation structure since we have repeated measures
  data = ChickWeight)
```

Let's see which model (basic or the one with auto-correlative structure) performs better:

```
anova(MultiLin_Base, MultiLin_Mod) # second model is better
```

##	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
##	MultiLin_Base	1 10	5487	5530	-2734			
##	MultiLin_Mod	2 11	4457	4505	-2217	1 vs 2	1032	<.0001

We clearly prefer the more sophisticated, auto-correlative model and want to see which of its parameters are informative:

```
anova(MultiLin_Mod) # all parameters should be kept
```

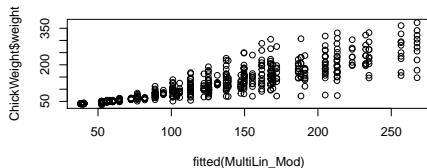
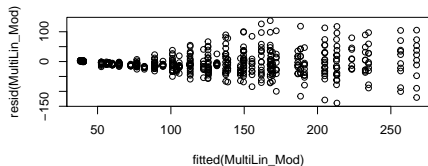
##		numDF	denDF	F-value	p-value
##	(Intercept)	1	524	485.8	<.0001
##	Time	1	524	1436.8	<.0001
##	Diet	3	46	3.5	0.0219
##	Time:Diet	3	524	24.9	<.0001

We keep all parameters. Although the inclusion of `Diet` is not significant, the interaction of `Diet` and `Time` is, therefore, both `Time` and `Diet` need to stay irrespective of their significance.

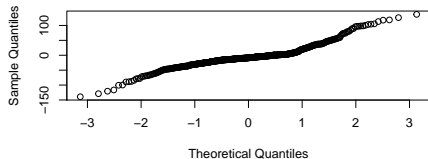


# Example - Assessing the Model

```
par(mfrow=c(2,2))
plot(fitted(MultiLin_Mod), resid(MultiLin_Mod)) # values around 0 -> good, fanning out -> bad
plot(fitted(MultiLin_Mod), ChickWeight$weight) # pattern fuzzy, but linear -> good
qqnorm(resid(MultiLin_Mod)) # residuals are not normal distributed -> bad
```



Normal Q-Q Plot



# Example - Model Output

```
summary(MultiLin_Mod)
```

```
## Linear mixed-effects model fit by REML
## Data: ChickWeight
##      AIC   BIC logLik
##  4457 4505  -2217
##
## Random effects:
## Formula: ~+1 | Chick
##      (Intercept) Residual
## StdDev:    0.006581    42.46
##
## Correlation Structure: AR(1)
## Formula: ~1 | Chick
## Parameter estimate(s):
##      Phi
## 0.9706
## Fixed effects: weight ~ Time * Diet
##              Value Std.Error DF t-value p-value
## (Intercept) 40.42      9.487 524   4.260 0.0000
## Time         6.06      0.350 524  17.347 0.0000
## Diet2        -0.88     16.430 46   -0.054 0.9575
## Diet3        -2.19     16.430 46   -0.133 0.8944
## Diet4        -1.01     16.431 46   -0.062 0.9512
## Time:Diet2    2.21      0.589 524   3.749 0.0002
## Time:Diet3    4.86      0.589 524   8.250 0.0000
## Time:Diet4    3.13      0.592 524   5.297 0.0000
## Correlation:
##      (Intr) Time    Diet2  Diet3  Diet4  Tm:Dt2 Tm:Dt3
## Time      -0.358
```

## Example - Model Output Explained

Variance between chicks (0.006581) is smaller than residual variance (42.46).  
This is good!

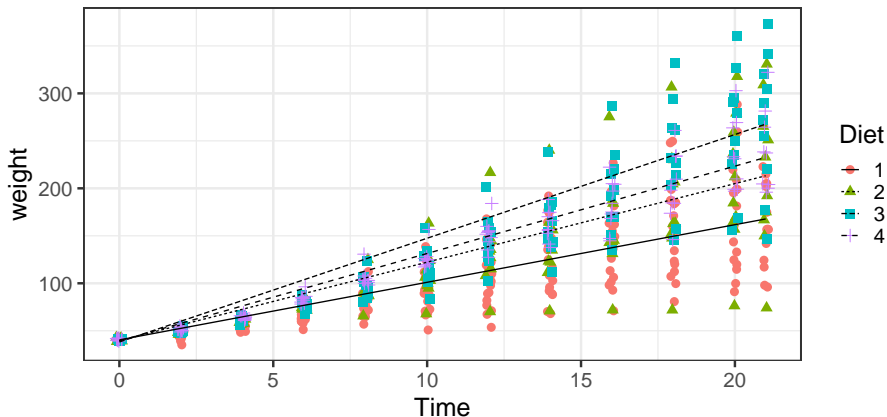
A chick is expected to have a `weight` of 40.4165 at `Time = 0` and `Diet = 1`.  
Per time-step, `weight` is expected to increase by 6.0637.

Mean chick `weight` is different to (read: “Diet1 weights are smaller by”) `Diet = 1` by -0.8794, -2.1932, and -1.0109 for `Diet = 2`, `Diet = 3`, and `Diet = 4` respectively (take note that these differences aren’t statistically significant).

For `Diet = 2`, `weight` of chicks increases, on average, by 2.2077 units more per one-unit increase in time when compared to `Diet = 1`. The same logic applies to `Diet = 3`, and `Diet = 4`.

# Example - Model Output Visualised

```
library(ggplot2)
ChickWeight$fit <- predict(MultiLin_Mod, level=0)
ggplot(ChickWeight, aes(Time, weight)) +
  geom_jitter(aes(colour=Diet, shape=Diet), width=0.1, size=3) +
  theme_bw(base_size=20) +
  geom_line(aes(y=fit, lty=Diet))
```



# Purpose & Assumptions

## Generalized Linear Models

`glm()` in base R

*Purpose:* Identify whether and how variables are related.

■ Down to *Study-Design*:

■ Variable values are **independent** (not paired)

*Assumptions:* ■ Need for *Post-Hoc Tests*:

■ Absence of **influential outliers**

■ Response and Predictor are related in a **linear** fashion

→ Allow for **non-normal** distributions and **heteroscedasticity**

# Purpose & Assumptions

## Generalized Linear Models

`glm()` in base R

*Purpose:* Identify whether and how variables are related.

■ Down to *Study-Design*:

■ Variable values are **independent** (not paired)

*Assumptions:* ■ Need for *Post-Hoc Tests*:

■ Absence of **influential outliers**

■ Response and Predictor are related in a **linear** fashion

→ Allow for **non-normal** distributions and **heteroscedasticity**

# Linear Predictor, Link Function, and Variance Function

## Components of a Generalized Linear Model:

**1 Linear predictor** e.g.:  $y_i = \alpha + \beta_1 x_i$

**2 Link function**  $g(\hat{y}_i) = y_i$

Relationship between predictor value and estimated value.

**3 Variance function**  $var(y_i) = \phi V_i(\bar{x})$

Variance depends on predictor mean, dispersion parameter  $\phi$  is a constant

Which **combinations** of components do I use?

Error	Link function	Variance function	Typical type of data
normal	identity	1 (constant)	Textbook examples
Poisson	$\log$	$var = \bar{x}$	Count data
binomial	$\text{logit}, \log(\bar{x}/(1 - \bar{x}))$	$var = \bar{x}(1 - \bar{x})/n$	Binary data

# Linear Predictor, Link Function, and Variance Function

## Components of a Generalized Linear Model:

**1 Linear predictor** e.g.:  $y_i = \alpha + \beta_1 x_i$

**2 Link function**  $g(\hat{y}_i) = y_i$

Relationship between predictor value and estimated value.

**3 Variance function**  $var(y_i) = \phi V_i(\bar{x})$

Variance depends on predictor mean, dispersion parameter  $\phi$  is a constant

## Which **combinations** of components do I use?

Error	Link function	Variance function	Typical type of data
normal	identity	1 (constant)	Textbook examples
Poisson	$\log$	$var = \bar{x}$	Count data
binomial	$\text{logit}, \log(\bar{x}/(1 - \bar{x}))$	$var = \bar{x}(1 - \bar{x})/n$	Binary data



# Example - The Data

```
library(AER)
data("PhDPublications")
head(PhDPublications)
```

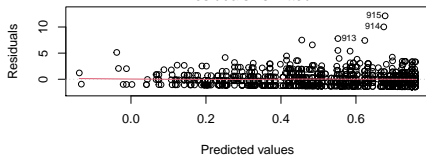
```
##   articles gender married kids prestige mentor
## 1         0   male      yes    0      2.52      7
## 2         0 female      no     0      2.05      6
## 3         0 female      no     0      3.75      6
## 4         0   male      yes    1      1.18      3
## 5         0 female      no     0      3.75     26
## 6         0 female      yes    2      3.59      2
```

→ Let's see if we can predict number of *articles* published by a PhD student as a result of their *gender*, *prestige*, and number if *kids*.

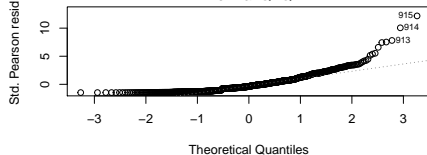
# Example - The Model

```
GeneralLin_Mod <- glm(articles ~ prestige*gender + kids, family = poisson, data = PhDPublications)
par(mfrow=c(2,2))
plot(GeneralLin_Mod)
```

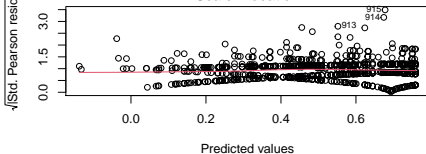
Residuals vs Fitted



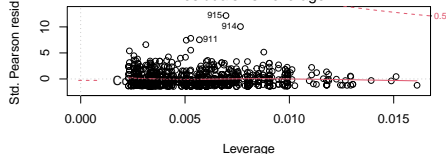
Normal Q-Q



Scale-Location



Residuals vs Leverage



# Example - Model Output

```
summary(GeneralLin_Mod)
```

```
##
## Call:
## glm(formula = articles ~ prestige * gender + kids, family = poisson,
##      data = PhDPublications)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.067  -1.606  -0.396   0.572   7.216
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.6228    0.1128   5.52 3.4e-08 ***
## prestige          0.0295    0.0330   0.90 0.37035
## genderfemale     -0.6993    0.1803  -3.88 0.00011 ***
## kids             -0.1258    0.0361  -3.49 0.00049 ***
## prestige:genderfemale 0.1265    0.0536   2.36 0.01828 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1766.9  on 910  degrees of freedom
## AIC: 3445
##
## Number of Fisher Scoring iterations: 5
```

## Example - Model Output Explained

According to this model, a male PhD student without kids and 0 prestige produces an estimated amount of 0.6228 papers during their PhD studies.

For an increase in prestige by 1, the number of papers increases by 0.0295 (this is not statistically significant). Unfortunately, being a women brings the average publication number down by -0.6993. Having kids also has an adverse effect of reducing the publication number by -0.1258 per kid. Lastly, an increase in prestige has more beneficial impacts on female PhD students as it increases their publication number per one-increase in prestige by 0.1265 from the baseline.

Do keep in mind that there are some issues that this model hasn't addressed yet. Especially, something we call **zero-inflation** of our response variable data as well as some predictor variable data.

## 1 The Basics

- Correlation Tests
- Regression Models
- Least Squares vs. Maximum Likelihood

## 2 Methods & Models

- Single Linear Regression
- Mixed Effect Models
- Generalised Linear Models

## 3 Choosing the Right Method

# Choices, Choices, Choices...

- **Linear Model** `lm`. When all **assumptions are met** (i.e.: homoscedasticity, normality, independence).
- **Linear Mixed Effect Model** `lme`. When the assumption of **independence is violated**.
- **Generalized Linear Model** `glm`. When the assumptions of **homoscedasticity** and **normality** are **violated**.
- **Generalized Linear Mixed Effect Model** `glmmPQL` from MASS, or `glmer` from `lme4`. When the assumptions of **homoscedasticity**, **normality**, and **independence** are **violated**.