REGRESSIONS

Correlations for the Advanced?

Erik Kusch

erik.kusch@au.dk

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

01/04/2020

1 The Basics
- Correlation Tests
- Regression Models
- Least Squares vs. Maximum Likelihood

2 Methods & Models
- Single Linear Regression
- Mixed Effect Models
- Generalised Linear Models

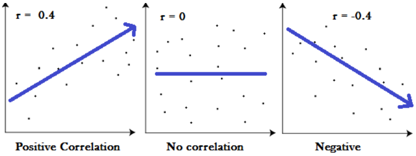
3 Choosing the Right Method

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)



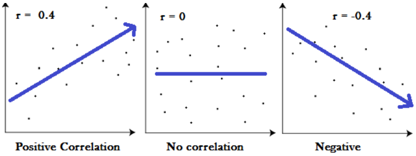
→ Get a feeling for it on Guess The Correlation.

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)



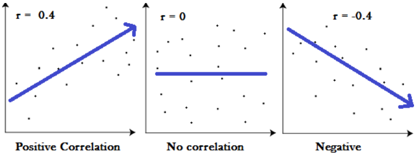
→ Get a feeling for it on Guess The Correlation.

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:

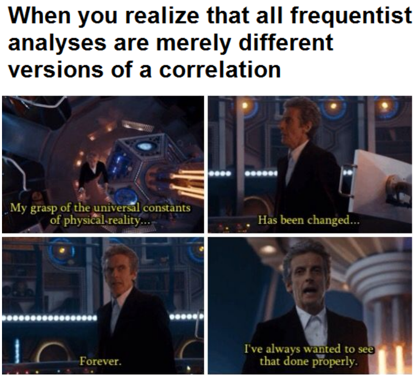
- Contingency Coefficient
- Kendall's Tau
- Spearman Correlation
- Pearson Correlation
- Cramer's V
- ANalysis Of VAriance (ANOVA)
- ...

Types of Correlations

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

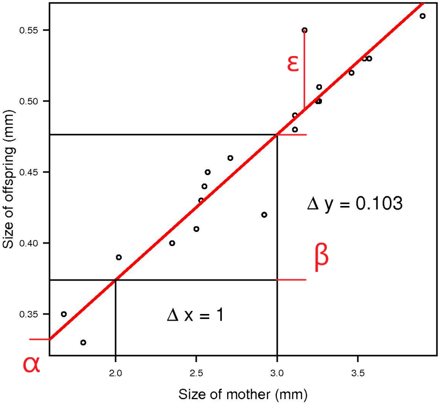
Prominent correlation tests include:

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



Terminology

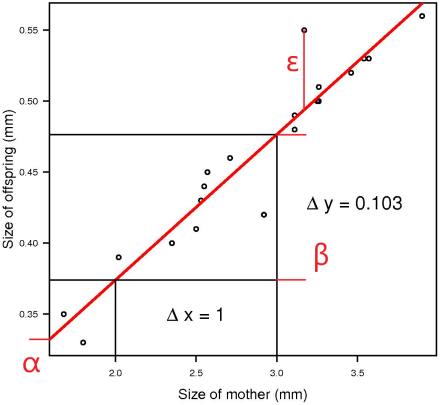
- α - The **Intercept**. The value of y when $x = 0$ (also referred to as β_0).
- β_i - The **Correlation Coefficient**. The increase in y for a one-unit increase in dependent variable i (usually, x if only one dependent variable).
- ϵ - The **Random Error**. The deviation of data points from the regression line. Usually assumed to follow $\epsilon \sim N(0, \sigma^2)$



Modified after Knut Helge Jensen.

Terminology

- α - The **Intercept**. The value of y when $x = 0$ (also referred to as β_0).
- β_i - The **Correlation Coefficient**. The increase in y for a one-unit increase in dependent variable i (usually, x if only one dependent variable).
- ϵ - 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.

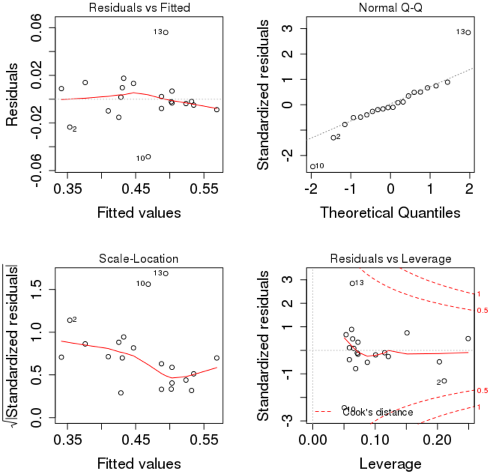
Assumptions in Theory

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

- **Residuals vs. Fitted values**
  Non-linear patterns identify a non-linear relationship between dependent and independent variables.
- **Normal Q-Q plot**
  Non-normal distribution of residuals shows that the assumption of $\epsilon \sim N(0, \sigma^2)$ is violated.
- **Scale Location**
  Non-constant variance identifies show that the assumption of homoscedasticity (invoked by least squares fitting).
- **Residuals vs. Leverage**
  A non-zero trend identifies the presence of influential outliers.

Assumptions in R

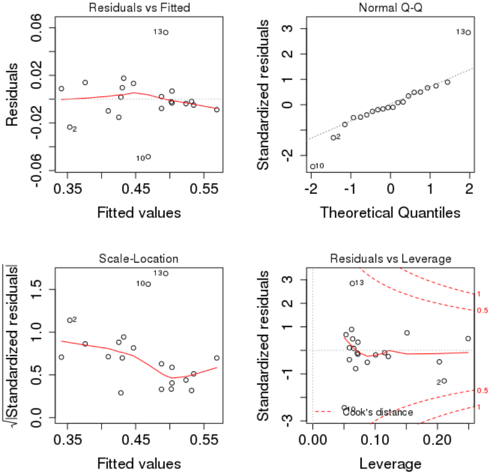
- Simply type ‘plot(...)‘ with ‘...‘ denoting your regression model.
- You can also target individual plots by writing:
  - ‘plot(..., 1)‘ for Residuals vs. Fitted values
  - ‘plot(..., 2)‘ for Normal Q-Q plot
  - ‘plot(..., 3)‘ for Scale Location
  - ‘plot(..., 4)‘ for Residuals vs. Leverage



By Knut Helge Jensen.

Assumptions in R

- Simply type ‘plot(...)‘ with ‘...‘ denoting your regression model.
- You can also target individual plots by writing:
  - ‘plot(..., 1)‘ for Residuals vs. Fitted values
  - ‘plot(..., 2)‘ for Normal Q-Q plot
  - ‘plot(..., 3)‘ for Scale Location
  - ‘plot(..., 4)‘ for Residuals vs. 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 $i = 1, ..., n$.

Minimize:

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

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 β_0 and β_1 such that the predicted probability $\hat{Pr}(x_j)$ corresponds to the observed response variable status.

Maximize:

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

Purpose & Assumptions

Single linear regression

`lm()` in base `R`

Purpose: Identify whether and how two variables are related.

Assumptions:

- Down to *Study-Design*:
  - Predictor variable is continuous (ratio or interval scale)
  - Response variable is continuous (ratio or interval scale)
  - Variable values are **independent** (not paired)
- 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.

Assumptions:

- Down to *Study-Design*:
  - Predictor variable is continuous (ratio or interval scale)
  - Response variable is continuous (ratio or interval scale)
  - Variable values are **independent** (not paired)
- 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

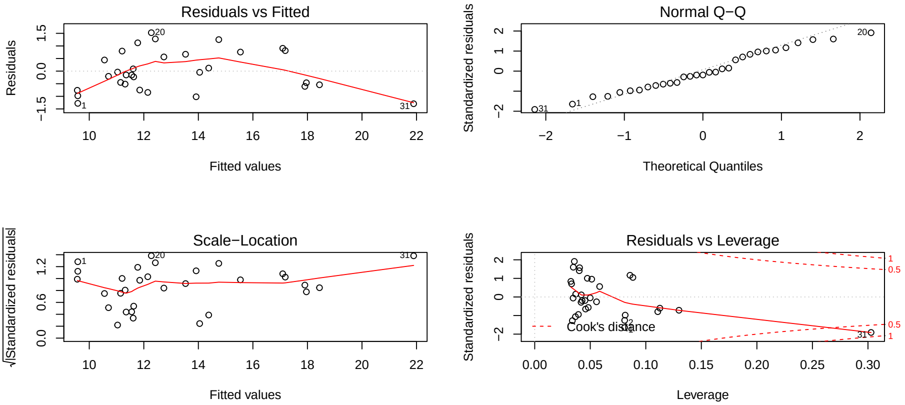
```r
# measures of Diameter (labelled as Girth), Height, and Volume of Timber
data("trees")
head(trees)
```

```
##   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
```

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

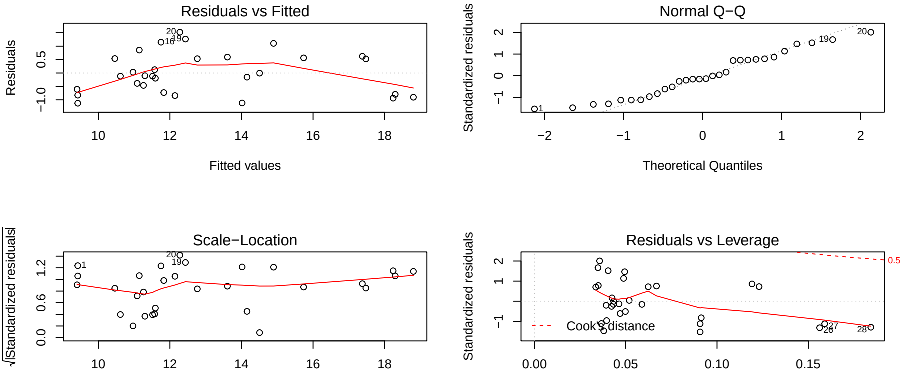
Example - The Model

```r
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 otulier in row 31
SingleLin_Mod <- with(trees, lm(Girth ~ Volume))
par(mfrow=c(2,2))
plot(SingleLin_Mod)
```



Example - Model Output

```r
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.

Assumptions:

- Down to *Study-Design*:
  - Predictor variable is continuous (ratio or interval scale)
  - Response variables are continuous (ratio or interval scale) and/or categorical (metric or ordinal scale)
- 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.

Assumptions:

- Down to *Study-Design*:
  - Predictor variable is continuous (ratio or interval scale)
  - Response variables are continuous (ratio or interval scale) and/or categorical (metric or ordinal scale)
- 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

```r
# 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

```r
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.

```r
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:

```r
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:

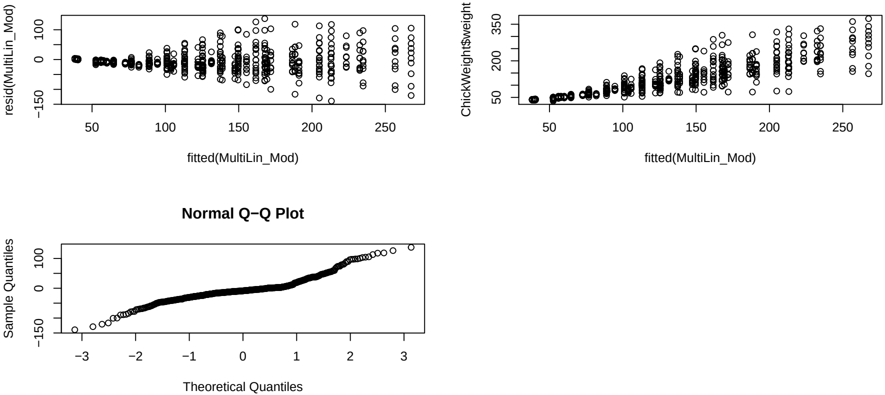
```r
anova(MultiLin_Mod) # all parameters should be kept
```

```
##             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
plot(fitted(MultiLin_Mod), ChickWeight$weight) # pattern fuzzy, but linear -> good
qqnorm(resid(MultiLin_Mod)) # residuals are not normal dsitributed -> bad
```



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) than residual variance (42.46). This is good!

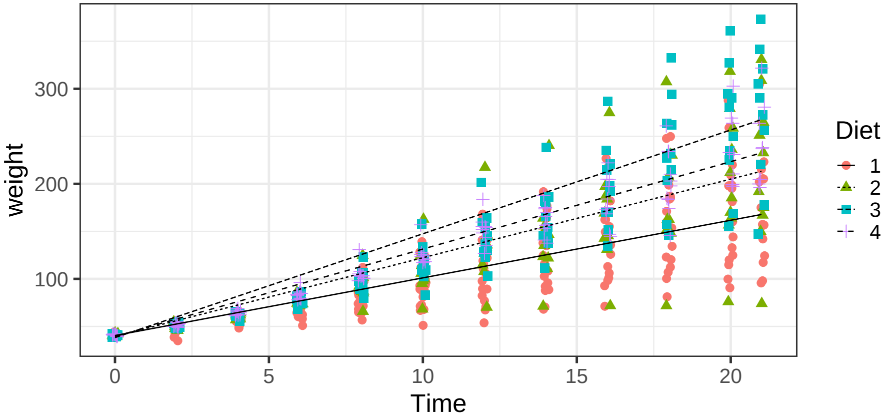
A chick is expected to have a `weight`of 40.4165 at $\texttt{Time} = 0$ and $\texttt{Diet} = 1$. Per time-step, `weight` is expected to increase by 6.0637.

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

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

Example - Model Output Visualised

```r
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.

Assumptions:

- Down to *Study-Design*:
  - Variable values are **independent** (not paired)
- 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.

Assumptions:

- Down to *Study-Design*:
  - Variable values are **independent** (not paired)
- 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** $var(y_i) = \phi V_i(\overline{x})$
   Variance depends on predictor mean, dispersion parameter ϕ 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	$var = \overline{x}$	Count data
binomial	logit, $log(\overline{x}/(1-\overline{x}))$	$var = \overline{x}(1-\overline{x})/n$	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** $var(y_i) = \phi V_i(\overline{x})$
   Variance depends on predictor mean, dispersion parameter ϕ 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	$var = \overline{x}$	Count data
binomial	logit, $log(\overline{x}/(1-\overline{x}))$	$var = \overline{x}(1-\overline{x})/n$	Binary data

Example - The Data

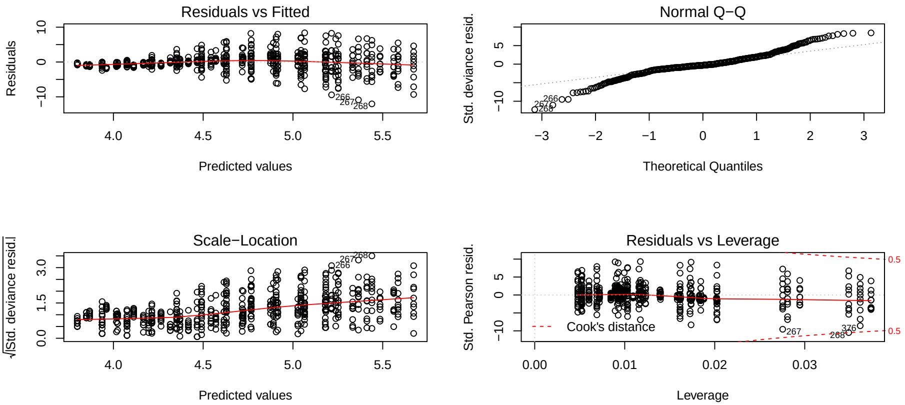
```r
data("ChickWeight")
head(ChickWeight)
```

```
## Grouped Data: weight ~ Time | Chick
##   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 reassess our earlier analysis of chick **weight** as a function of **time** and **diet**. This time, we will forego the random effect of chicks since that would create a generalized linear mixed effect model.

Example - The Model

```
GeneralLin_Mod <- glm(weight ~ Time*Diet, family = poisson, data = ChickWeight)
par(mfrow=c(2,2))
plot(GeneralLin_Mod)
```



Example - Model Output

```r
summary(GeneralLin_Mod)
```

```
##
## Call:
## glm(formula = weight ~ Time * Diet, family = poisson, data = ChickWeight)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -12.028   -1.198   -0.307    1.169    8.376
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.79881    0.01565  242.75  < 2e-16 ***
## Time         0.06922    0.00105   66.03  < 2e-16 ***
## Diet2        0.06859    0.02584    2.65  0.00794 **
## Diet3        0.05064    0.02539    1.99  0.04615 *
## Diet4        0.16160    0.02511    6.44  1.2e-10 ***
## Time:Diet2   0.00565    0.00170    3.33  0.00086 ***
## Time:Diet3   0.01758    0.00164   10.69  < 2e-16 ***
## Time:Diet4   0.00701    0.00166    4.23  2.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22444.4  on 577  degrees of freedom
## Residual deviance:  3996.3  on 570  degrees of freedom
## AIC: 7755
##
## Number of Fisher Scoring iterations: 4
```

Example - Model Output Explained

A chick is now expected to have a `weight`of 3.7988 at `Time` $= 0$ and `Diet` $= 1$. Per time-step, `weight` is expected to increase by 0.0692.

Mean chick `weight` of `Diet` $= 1$ is smaller by 0.0686, 0.0506, and 0.1616 for `Diet` $= 2$, `Diet` $= 3$, and `Diet` $= 4$ respectively (these differences are now significant).

`weight` of chicks increases, on average, increases by 0.0056 units more per one-unit increase in time when compared to `Diet` $= 1$. The same logic applies to `Diet` $= 3$, and `Diet` $= 4$.

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**.