STAT9006: Correlation and Regression with *R* Part I

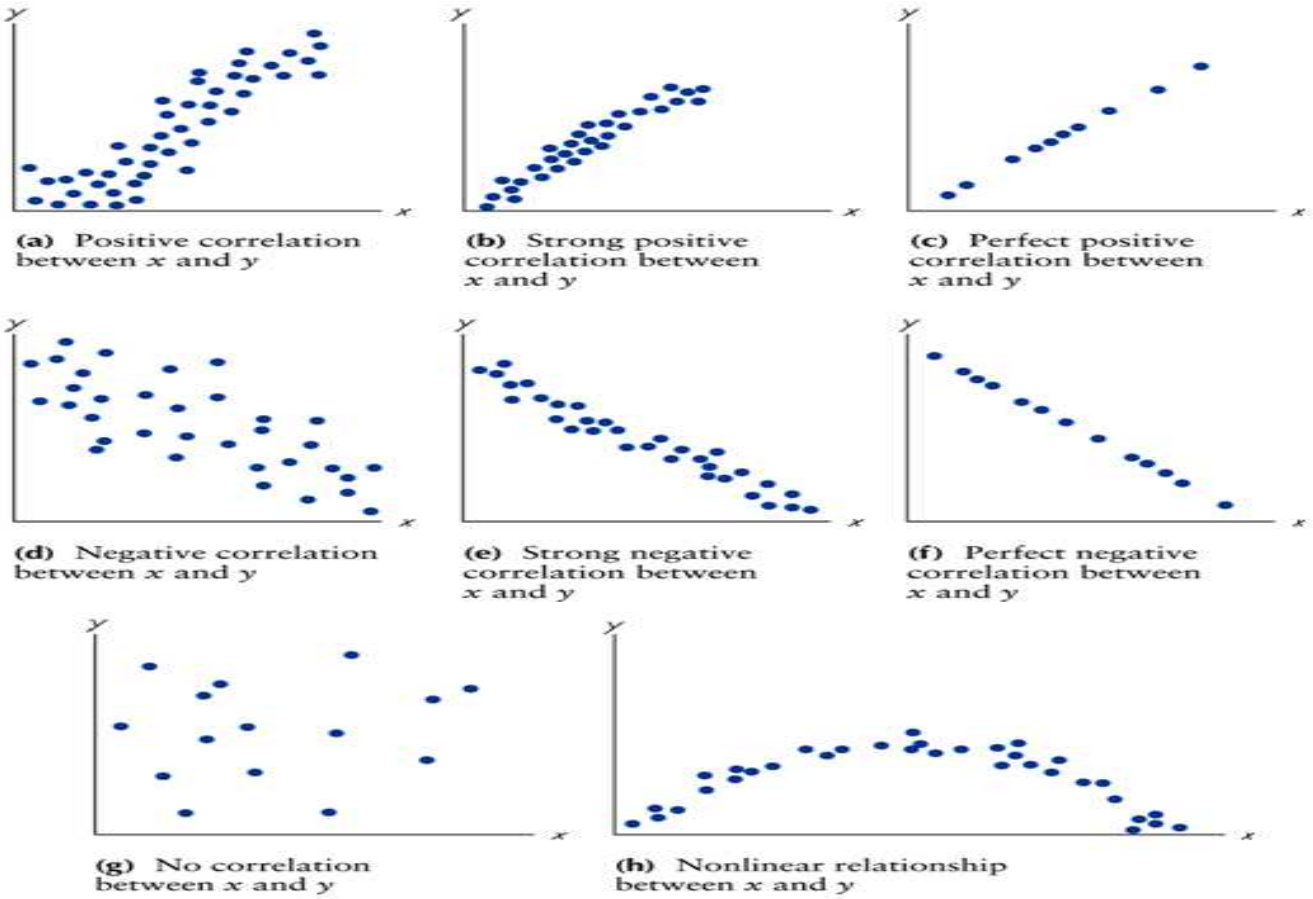
Correlation and regression

1. Correlation and regression
   - Structure
2. Example
   - Format and Explore
   - Scatterplot
   - Correlation
   - Regression
   - Residuals
3. Exercise

Scatterplot

- A scatterplot is a **two dimensional plot** showing the (X,Y) value for each observation.
- It is used to determine whether there is any **pronounced relationship** and, if so, whether the relationship can be treated as approximately linear.
- Another use of the scatterplot is the **detection of outliers**, (which could affect normality).
- Y is usually the **response variable**. The response variable is the variable for which you **want to explain the variation**.
- X is usually the **explanatory variable**. The explanatory variable is the variable **used to explain variation** in the response variable.

Scatterplots



(a) Positive correlation between x and y

(b) Strong positive correlation between x and y

(c) Perfect positive correlation between x and y

(d) Negative correlation between x and y

(e) Strong negative correlation between x and y

(f) Perfect negative correlation between x and y

(g) No correlation between x and y

(h) Nonlinear relationship between x and y

Scatterplots

- Figure (a)-(c) are examples of **positive linear** relationships between an explanatory variable and a response variable. This means that as the as x increases in value, y increases and tends to change systematically in a positive straight line.
- Figure (d)-(f) are examples of **negative linear** relationships between x and y. This means that as the as x increases in value, y decreases and tends to change systematically in a straight line.
- Figure (g)-(h) are examples of where **no linear** relationship exists.

Correlation

- The word correlation is used to describe the relationship between two or more variables.
- If the sample is **normally distributed** then the **Pearson** correlation coefficient is used.
- If the sample is **not normally distributed** then **Spearman's rho** coefficient is used.

Properties of the linear correlation coefficient

1. $-1 \leq r \leq 1$,
2. $r = \pm 1$ represents a perfect positive/negative linear correlation between the variables.
3. $r = 0$ indicates little or no linear relationship.
4. The more r differs from 0, the stronger the linear relationship between the two variables.
5. The sign of r indicates the direction of the relationship
6. r is not affected by the choice of x and y. Interchange x and y and the value of r will not change.
7. r measures strength of a linear relationship.
8. The value of r^2 (coefficient of determination) is the proportion of the variation in y that is explained by the linear relationship between x and y.

Errors in correlation

Three points to be aware off when determining linear correlation are:

- **Causation** - It is wrong to conclude that correlation implies causality.
- **Averages** - may suppress individual variation and may inflate the correlation coefficient.
- **Linearity** - There may be some other type of relationship between x and y even when there is no significant linear correlation.

Regression

- Regression analysis enables you to **predict** the value of one variable given the value of another.
- It gives you an **equation** that uses one variable to help explain variation in another.
- If you were to attempt to draw a curve through the points on a scatterplot that **models the relationships between two variables**, there are several curves you could draw.
- Regression analysis fits the curve that **best represents the relationship** - i.e., the curve that has the minimum distance from each of the points.
- The distance from the line to any of the points is called the **residual**. The 'best fit line minimises the sum of the residuals.

Guidelines for using the regression equation

Overall, the derivation of the regression line is used for prediction - it enables prediction of the value of one variable given the value of another. Regression analysis yields an equation that uses one variable to help explain variation in another variable.
Guidelines for using the regression equation are:

- If there is **no significant linear correlation**, do not use the regression equation to make predictions.
- When using the regression equation for predictions, **stay within the scope** of the available sample data.
- A regression equation based on **old data** is not necessarily valid now.
- Do not make predictions about a **population** that is different from the population from which the sample data was drawn.

Example 01

A fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the residence and the nearest fire station. The study is to be conducted in a large suburb of a major city using a sample of 15 recent fires. For each fire, the amount of damage in thousands of euro and the distance between the location of the fire and the nearest station are recorded. (Import *Example01*.xlsx)

Format

```
# FIRST format and explore the data
# Step 01: format the data
# change Fire Damage names to make script more manageable (if desired)
names(FD)<-c("Distance","Damage")

### append a case number ... from Workshop 03
(n<-dim(FD)[1]) # sample size
CaseNum<-seq(1:n)
library(dplyr)
FD<-mutate(FD,CaseNum)

#### ordering variables (if desired)
(cn<-dim(FD)[2])
FD<-FD[,c(cn,1:cn-1)]

# Step 02: check properties
prop<-c() # setting up a vector/variable
for(i in 2:cn){
  prop[i]<-is.numeric(FD[[i]])
}
prop
```

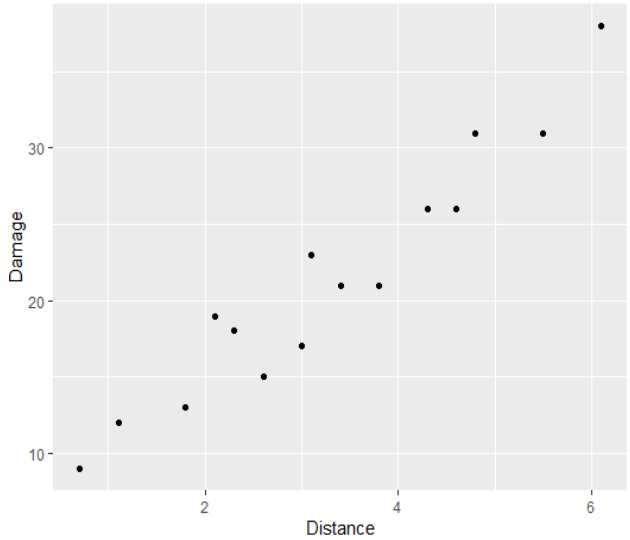
Explore

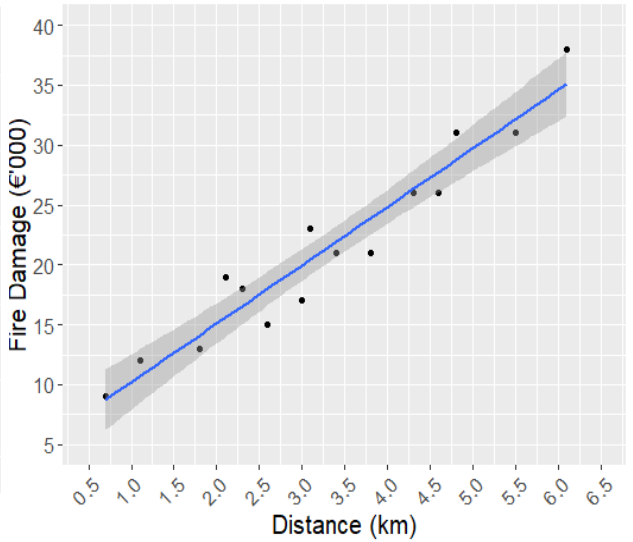
```
> # Step 03: Explore the data
> # numerical descriptive statistics
> # stacking data is easiest way to manage this
> library(tidyr)
> Long<-gather(FD,Variable,Value,2:3)
> Stats<-Long %>% group_by(Variable) %>% summarise("Sample size"=n(),Mean = mean(Value),
+                                                  "Standard deviation"=sd(Value),
+                                                          Median = median(Value),
+                                                  "1st quartile"=quantile(Value, 0.25),
+                                                          "3rd quartile"=quantile(Value, 0.75),
+                                                  Min=min(Value), Max=max(Value))
> t(Stats)
                    [,1]       [,2]
Variable            "Damage"   "Distance"
Sample size         "15"       "15"
Mean                "21.33333" " 3.28000"
Standard deviation  "7.997023" "1.576252"
Median              "21.0"     " 3.1"
1st quartile        "16.0"     " 2.2"
3rd quartile        "26.00"    " 4.45"
Min                 "9.0"      "0.7"
Max                 "38.0"     " 6.1"
```

Scatterplot

Input *Distance* as the explanatory variable, x, and the *Damage* as the response variable, y, respectively

```
# SECOND scatterplot (using original dataframe)
library(ggplot2)
(g<-ggplot(FD,aes(x=Distance,y=Damage))+geom_point())
(g1<-g+labs(x="Distance (km)", y="Fire Damage (€'000)")+ #label axes
  coord_cartesian(xlim=c(0.5,6.5),ylim=c(5,40))+ #scale axes
    scale_x_continuous(breaks=seq(0.5,6.5,0.5))+
  scale_y_continuous(breaks=seq(5,40,5)))
(g2<-g1+theme(text = element_text(size=15), # increasing font size
              axis.text.x = element_text(angle = 45, hjust = 1))) # rotating angle of text
g2+geom_smooth(method=lm,se=T) # including regression line
```





Tests for correlation

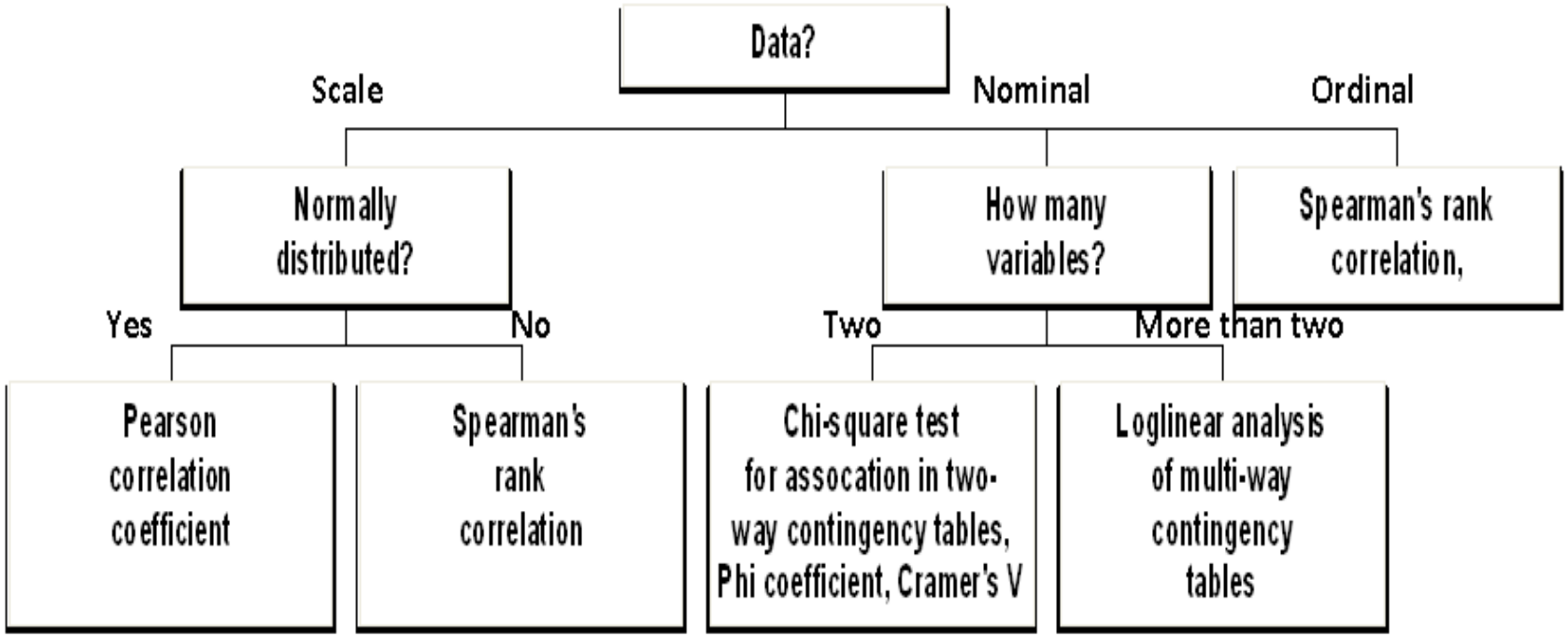
The word correlation is used to describe the relationship between two or more variables.

- H_0: **No correlation** exists between the two variables.
- H_1: **A correlation** exists between the two variables.

Remember if:

- If $p \leq \alpha$, then reject H_0 - i.e., reject the claim that no correlation exists between the two variables;
- If $p > \alpha$, then fail to reject H_0 - i.e., fail to reject the claim that no correlation exists between the two variables.

Selecting a test for correlation



Normally distributed data?

```
> # THIRD Correlation
> # Step 01: Tests of normality (return to long format)
> library(psych)
> Norm<-Long %>% group_by(Variable) %>% summarise("Sample size"=n(),Mean = mean(Value),
+                                                 Median = median(Value), Skewness=skew(Value),
+                                                 "Normally distributed"=ifelse(
+                                                   shapiro.test(Value)$p.value>0.05,"Yes","No"),
+                                                "p-value"=round(shapiro.test(Value)$p.value,4))
> t(Norm)
                    [,1]        [,2]
Variable            "Damage"    "Distance"
Sample size         "15"        "15"
Mean                "21.33333"  " 3.28000"
Median              "21.0"      " 3.1"
Skewness            "0.3899027" "0.1137152"
Normally distributed "Yes"      "Yes"
p-value             "0.9008"    "0.9862"
```

Example: Correlation Coefficient

If the data is normally distributed then use the Pearson correlation coefficient by, other the Spearman correlation coefficient:

```
> library(psych)
> # [,-1] excludes CaseNum
> (res1<-corr.test(FD[,-1])) #defaults to Pearson and displays correlation and p-value
Call:corr.test(x = FD[, -1])
Correlation matrix
         Distance Damage
Distance     1.00   0.96
Damage       0.96   1.00
Sample Size
[1] 15
Probability values (Entries above the diagonal are adjusted for multiple tests.)
         Distance Damage
Distance        0      0
Damage          0      0

 To see confidence intervals of the correlations, print with the short=FALSE option
> (res2<-corr.test(FD[,-1],method="spearman")) # specify non-parametric test, if required
Call:corr.test(x = FD[, -1], method = "spearman")
Correlation matrix
         Distance Damage
Distance     1.00   0.95
Damage       0.95   1.00
Sample Size
[1] 15
Probability values (Entries above the diagonal are adjusted for multiple tests.)
         Distance Damage
Distance        0      0
Damage          0      0
```

Effect size

The following table offers a rough guide to the classification of effect size in regression - i.e., the strength of the relationship.

Size of effect	Absolute value of r	r squared
Small	$0.1 \leq \lvert r \rvert < 0.3$	$0.01 \leq r^2 < 0.09$
Medium	$0.3 \leq \lvert r \rvert < 0.5$	$0.09 \leq r^2 < 0.25$
Large	$\lvert r \rvert \geq 0.5$	$r^2 \geq 0.25$

Linear regression analysis

If there is a linear correlation between the two variables then it is possible to determine a linear regression line that best fits the data.

```
> # FOURTH regression (if correlation is significant)
> fit<-lm(Damage~Distance, FD) #returns regression coefficients
> summary(fit) #returns more info

Call:
lm(formula = Damage ~ Distance, data = FD)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0178 -1.4635 -0.3067  1.8502  3.4201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3407     1.4064   3.797  0.00222 **
Distance      4.8758     0.3889  12.537 1.23e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.294 on 13 degrees of freedom
Multiple R-squared:  0.9236,    Adjusted R-squared:  0.9177
F-statistic: 157.2 on 1 and 13 DF,  p-value: 1.234e-08
```

- H_0: **No difference** exists between the slope of the regression line and an horizontal line (i.e., $\beta = 0$).
- H_1: **A difference** exists between the slope of the regression line and an horizontal line (i.e., $\beta \neq 0$).

Linear regression analysis

Types of output:

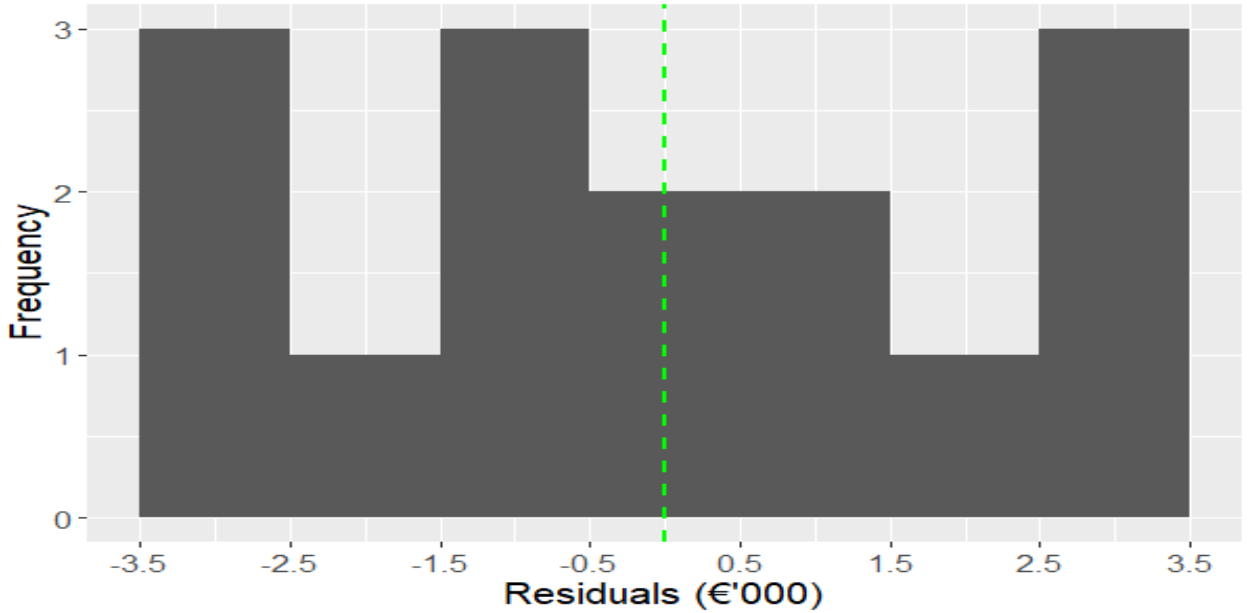
```
> coefficients(fit) # model coefficients/parameters
(Intercept)    Distance
   5.340693    4.875805
> confint(fit, level=0.95) # CIs for model coefficients/parameters
                2.5 %   97.5 %
(Intercept) 2.302300 8.379086
Distance    4.035602 5.716008
> round(fitted(fit),2) # predicted values
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15
21.92 14.12 27.77 16.56 20.46 32.16  8.75 19.97 18.02 26.31 15.58 10.70 35.08 28.74 23.87
> round(residuals(fit),2) # residuals
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15
-0.92 -1.12 -1.77  1.44  2.54 -1.16  0.25 -2.97 -3.02 -0.31  3.42  1.30  2.92  2.26 -2.87
```

Residuals: Normality

- **Histogram of Residuals** - generally used as a pointer towards the distribution of the residuals. What you are looking for is a bell-shaped curve. Because the shape can differ depending on the width of the intervals, the normal probability plot is a more reliable indicator.
- **Test of normality** - generally used to support the interpretation of the above mentioned plot.
- **Test of differences** - generally used to whether the residuals, on average, are different from 0.

Residuals: Histogram

```
# FIFTH RESIDUALS
# normal
range(residuals(fit)) # to have an idea of the spread of the data from # breaks in histogram
# fortify() converts fit to a dataframe
(n<-ggplot(fortify(fit), aes(x = .resid)) + geom_histogram(breaks=seq(-3.5,3.5,1)))
(final_n<-n+labs(x="Residuals (€'000)", y="Frequency")+
    coord_cartesian(xlim=c(-3.5,3.5))+scale_x_continuous(breaks=seq(-3.5,3.5,1))+
    geom_vline(xintercept=0, linetype="dashed", color = "green", size=1)+ #different from 0
    theme(text = element_text(size=15)))
```



Residuals: Test of Normality/Differences

```
> shapiro.test(residuals(fit)) #normality of residuals

        Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.93164, p-value = 0.2887

> t.test(residuals(fit),mu=0) # different from zero?

        One Sample t-test

data:  residuals(fit)
t = 1.9137e-16, df = 14, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.224031  1.224031
sample estimates:
   mean of x
1.092153e-16
```

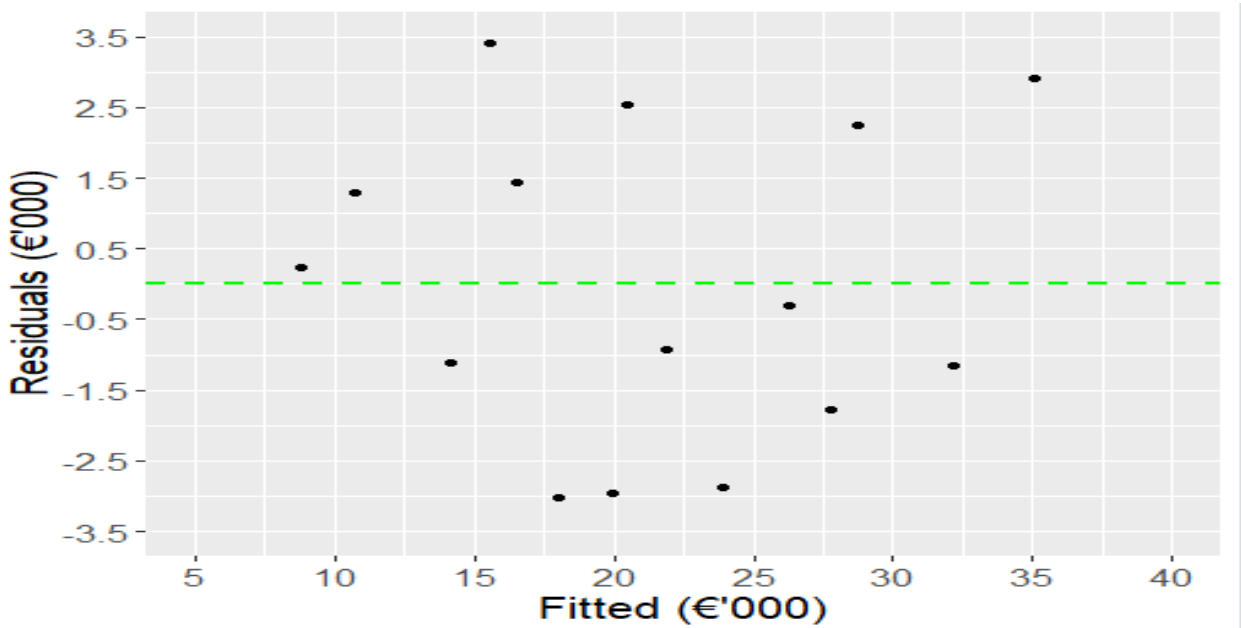
Residuals: Random

Plotting the residuals vs fitted values will reveal any correlated error terms or **heteroscedasticity**.

- **Residuals Versus Fits Plot** - used to check for constant variance. The data should have no pattern.
- **Durbin Watson test** - generally used to support the interpretation of the above mentioned plot.

Residuals: Scatterplot

```
# Random
# fortify() converts fit to a dataframe
(r<-ggplot(fortify(fit), aes(x = .fitted, y = .resid)) + geom_point()) #better looking
(final_r<-r+labs(x="Fitted (€'000)", y="Residuals (€'000)")+
  coord_cartesian(xlim=c(5,40),ylim=c(-3.5,3.5))+scale_x_continuous(breaks=seq(5,40,5))+
  scale_y_continuous(breaks=seq(-3.5,3.5,1))+
  geom_hline(yintercept=0, linetype="dashed", color = "green", size=1)+
    theme(text = element_text(size=15)))
```



Residuals: Durbin Watson test

```
> library(lmtest)
> dwtest(fit) #randomness of residuals

        Durbin-Watson test

data:  fit
DW = 1.361, p-value = 0.1126
alternative hypothesis: true autocorrelation is greater than 0
```

Exercise 01

A wood scientist wants to establish if there is a relationship between the adhesive strength of laminated wood and the dwell time in the press machine. A random sample of 7 different times and their corresponding adhesive strengths in pounds per square inch were recorded (*Exercise01*.xlsx). Using the steps outlined in the above slides:

1. Test if there is a statistically significant relationship between time and adhesive strength.
2. If a statistically significant relationship exists, then use regression analysis to model the relationship.
3. Confirm that the assumptions placed on the residuals of the model are satisfied.