STAT9006: Correlation and Regression with R Part II

CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

MATHEMATICS

1. Multiple regression
   - Introduction
   - Multi-collinearity
2. Example
   - Format and Explore
   - Scatterplots
   - Correlation
   - Multiple Linear Regression
   - Residuals
3. Exercise

Introduction

- The ultimate objective of a multiple regression analysis is to develop a model that will accurately predict a dependent variable (y) as a function of a **set of independent variables** $(x_1, x_2, x_3, x_4, ..., x_k)$.

- The most difficult part of regression analysis is **choosing the correct model** for a practical application

Structure

There is a basic step by step approach that you can use when performing a multiple regression analysis (after exploring the data for outliers and producing scatterplots/correlations):

1. Obtain a fitted prediction **model**;
2. Use the **adjusted** R^2 to determine how well the model fits the data;
3. Check the t-tests for the **partial regression coefficients** to see which ones are contributing significant information in the presence of others;
4. If you choose to compare several different models, use adjusted R^2 to compare their effectiveness;
5. Use **residual plots** to check for violation of the regression assumptions.

Multi-collinearity

The following are indicators that your regression model is likely to exhibit multi-collinearity during your analysis:

- The value of the R^2 **is large** indicating a good fit, but the individual **regression coefficients are insignificant**;
- The signs of the regression coefficients are not what you would have expected them to be - e.g., contributing negatively rather than positively;
- A matrix of correlations shows that some of the **independent (predictor) variables are highly correlated**;
- The **Variance Inflation Factor** (VIF) for some predictors is high.

Variance Inflation Factor (VIF)

#*vif*()

- A value of 1 means that the predictor is not correlated with other variables.
- The higher the value, the greater the correlation of the variable with other variables.
- Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.
- These numbers are just rules of thumb; in some contexts a VIF of 2 could be a great problem, whereas in straightforward predictive applications very high VIFs may be unproblematic.

Multi-collinearity

- If multi-collinearity exists in your model, you need to remove the variables that are causing this.
- The VIF can help guide you in the decision as to what variables should be removed.
- **Stepwise regression** can be a useful procedure to determine the best model. But, note, that stepwise regression removes variables that cause multi-collinearity and/or variables that are statistically insignificantly contributing to the model.
- **Partial Least Squares** (PLS) regression and **Principal Component Regression** (PCR) are who other methods than can be useful to reduce a large number of correlated variables to form a **smaller number of uncorrelated variables/factors** from a large set of data.

Types of Stepwise Regression

1. **Forward selection**, which starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant. This method often applies to where the number of samples n is inferior to the number of predictors.

2. **Backward selection** starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when with a model where all predictors are statistically significant. This method often requires that the number of samples n is larger than the number of predictors.

3. **Stepwise selection** is a combination of forward and backward selections. Starts with no predictors, then sequentially adds the most contributive predictors (like forward selection). After adding each new variable, removes any variables that no longer provide an improvement in the model fit (like backward selection). This method often applies to where the number of samples n is inferior to the number of predictors.

Partial Least Squares Regression

- Partial Least Squares (PLS) regression identifies new principal components that not only summarises the original predictors, but also that are **related to the response variable**.
- These components are then used to fit the regression model.
- PLS uses a **dimension reduction** strategy that is supervised by the response variable.
- PLS is convenient for data with **highly-correlated predictors**.
- The number of PCs used in PLS is generally chosen by **cross-validation**.
- Predictors and the response variables should be generally **standardised**, to make the variables comparable.

Principal Component Regression

- The principal component regression (PCR) first applies **Principal Component Analysis** on the data set to summarize the original predictor variables into few new variables also known as principal components (PCs).
- PCs are a linear combination of the original data.
- These PCs are then used to build the linear regression model.
- The number of principal components, to incorporate in the model, is chosen by **cross-validation** (cv).
- Note that, PCR is suitable when the data set contains **highly correlated predictors**.
- A possible **drawback** of PCR is that we have no guarantee that the selected principal components are associated with the response variable.

Example 02

Suppose the sales manager of a large automotive parts distributor wants to estimate the total *annual sales* for a region. On the basis of regional sales, the total sales for the company can also be estimated. Several factors appear to be related to sales, including the *number of retail outlets* in the region stocking the companys products, the *number of automobiles* in the region, and the total *personal income* for the first quarter of the year. In total five independent variables were finally selected as being the most important (according to the sales manager). The data was gathered for a recent year. The total annual sales were also recorded for each region for that year. This data is presented in *Example02*.xlsx. Using multiple regression, find the model that best fits the data.

Format

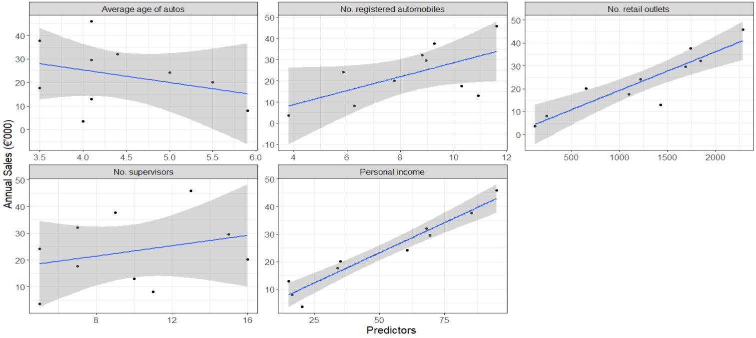
```
# FIRST format and explore the data
# Step 01: format the data
### append a case number ... from Workshop 03
(n<-dim(AS)[1]) # sample size
CaseNum<-seq(1:n)
library(dplyr)
AS<-mutate(AS,CaseNum)

#### ordering variables (if desired)
(cn<-dim(AS)[2])
AS<-AS[,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(AS[[i]])
}
prop
```

Explore

```
> # Step 03: Explore the data
> # numerical descriptive statistics
> # stacking data is easiest way to manage this
> library(tidyr)
> Long<-gather(AS,Variable,Value,2:7)
> 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]                  [,3]                         [,4]                [,5]              [,6]
Variable           "Annual sales" "Average age of autos" "No. registered automobiles" "No. retail outlets" "No. supervisors" "Personal income"
Sample size        "10"           "10"                  "10"                         "10"                "10"              "10"
Mean               "  23.1932"    "  4.4100"            "  8.3610"                   "1231.6000"         "  9.8000"        "  49.8900"
Standard deviation " 13.345386"   " 0.807534"           " 2.449519"                  "713.020835"        " 3.938415"       " 29.502973"
Median             "  22.156"     "  4.100"             "  8.885"                    "1324.000"          "  9.500"         "  47.800"
1st quartile       " 14.15175"    " 4.02500"            " 6.65250"                   "760.75000"         " 7.00000"        " 23.60000"
3rd quartile       "  31.44125"   "  4.85000"           " 10.05000"                  "1726.00000"        " 12.50000"       "  69.00000"
Min                "  3.611"      " 3.500"              " 3.810"                     "120.000"           " 5.000"          " 15.100"
Max                "  45.919"     "  5.900"             " 11.620"                    "2290.000"          " 16.000"         "  95.100"
```

Scatterplots

```
# SECOND scatterplot (using original dataframe)
library(tidyr)
library(ggplot2)
(gs<-AS[,-1] %>% gather(-`Annual sales`,key=var,value="value") %>%
    ggplot(aes(x=value,y=`Annual sales`))+geom_point()+
    facet_wrap(~ var, scales="free")+theme_bw())
gs+labs(x="Predictors",y="Annual Sales (€'000)")+
          geom_smooth(method="lm")+ # includes regression line
   theme(text = element_text(size=15))
```



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]                   [,3]                         [,4]                [,5]              [,6]
Variable             "Annual sales" "Average age of autos" "No. registered automobiles" "No. retail outlets" "No. supervisors" "Personal income"
Sample size          "10"           "10"                   "10"                         "10"                "10"              "10"
Mean                 "  23.1932"    "   4.4100"            "   8.3610"                  "1231.6000"         "   9.8000"       "  49.8900"
Median               "  22.156"     "   4.100"             "   8.885"                   "1324.000"          "   9.500"        "  47.800"
Skewness             " 0.1493850"   " 0.5968304"           "-0.4033955"                 "-0.2471078"        " 0.2420068"      " 0.1483067"
Normally distributed "Yes"          "Yes"                  "Yes"                        "Yes"               "Yes"             "Yes"
p-value              "0.9849"       "0.1715"               "0.7836"                     "0.6958"            "0.4870"          "0.2762"
```

Correlation coefficient matrix

```
> # Step 02: correlation test
> library(psych)
> # [,-1] excludes CaseNum
> (res1<-corr.test(AS[,-1])) # defaults to Pearson
Call:corr.test(x = AS[, -1])
Correlation matrix
                          Annual sales No. retail outlets No. registered automobiles Personal income Average age of autos No. supervisors
Annual sales                      1.00               0.90                       0.60            0.96                -0.32            0.29
No. retail outlets                0.90               1.00                       0.78            0.82                -0.49            0.18
No. registered automobiles        0.60               0.78                       1.00            0.41                -0.45            0.40
Personal income                   0.96               0.82                       0.41            1.00                -0.35            0.15
Average age of autos             -0.32              -0.49                      -0.45           -0.35                 1.00            0.29
No. supervisors                   0.29               0.18                       0.40            0.15                 0.29            1.00
Sample Size
[1] 10
Probability values (Entries above the diagonal are adjusted for multiple tests.)
                          Annual sales No. retail outlets No. registered automobiles Personal income Average age of autos No. supervisors
Annual sales                      0.00               0.01                       0.70            0.00                 1.00               1
No. retail outlets                0.00               0.00                       0.10            0.04                 1.00               1
No. registered automobiles        0.06               0.01                       0.00            1.00                 1.00               1
Personal income                   0.00               0.00                       0.24            0.00                 1.00               1
Average age of autos              0.36               0.15                       0.20            0.32                 0.00               1
No. supervisors                   0.42               0.61                       0.26            0.67                 0.42               0

 To see confidence intervals of the correlations, print with the short=FALSE option
```

Formatting correlation coefficient matrix

```
> # correlation of response variable with predictors
> (out<-rbind(t(round(res1$r[1,],4)),t(round(res1$p[1,],4))))
     Annual sales No. retail outlets No. registered automobiles Personal income Average age of autos No. supervisors
[1,]            1             0.8994                     0.6048          0.9645              -0.3227          0.2858
[2,]            0             0.0055                     0.7039          0.0001               1.0000          1.0000
>
> # investigate if multi-collinearity present
> cm<-corr.test(AS[,3:7])$r # focusses on predictors only
> library(Matrix)
> round(tril(cm),4) # lower triangular
5 x 5 Matrix of class "dtrMatrix"
                           No. retail outlets No. registered automobiles Personal income Average age of autos No. supervisors
No. retail outlets                     1.0000                          .               .                    .               .
No. registered automobiles             0.7752                     1.0000               .                    .               .
Personal income                        0.8249                     0.4088          1.0000                    .               .
Average age of autos                  -0.4894                    -0.4465         -0.3495               1.0000               .
No. supervisors                        0.1833                     0.3951          0.1546               0.2907          1.0000
```

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	$\lvert r \rvert \geq 0.5$	$r^2 \geq 0.25$

Multiple linear regression analysis

- 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., $\beta \neq 0$).

```
> # FOURTH regression (if correlation is significant)
> # for ease of typing make name of variables shorter
> names(AS)<-c("CaseNum","Sales","Outlets","RA","PI","AA","Supervisors")
>
> # Full model
> fit1<-lm(Sales~Outlets+RA+PI+AA+Supervisors,AS) #lm(Sales ~., data = AS[,-1])
> summary(fit1)

Call:
lm(formula = Sales ~ Outlets + RA + PI + AA + Supervisors, data = AS)

Residuals:
       1        2        3        4        5        6        7        8        9       10 
 0.51432 -0.45018  0.92579  0.47761 -0.69260 -0.07109 -1.49579 -1.30332  1.72442  0.37083 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.967e+01  5.422e+00  -3.628 0.022195 *  
Outlets     -6.286e-04  2.638e-03  -0.238 0.823391    
RA           1.740e+00  5.530e-01   3.146 0.034638 *  
PI           4.099e-01  4.385e-02   9.348 0.000729 ***
AA           2.036e+00  8.779e-01   2.319 0.081238 .  
Supervisors -3.445e-02  1.880e-01  -0.183 0.863534    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.507 on 4 degrees of freedom
Multiple R-squared:  0.9943,	Adjusted R-squared:  0.9872 
F-statistic: 140.4 on 5 and 4 DF,  p-value: 0.0001397
```

Multi-collinearity in the model

```
> # Check for multi-collinearity
> library(faraway)
> (v1<-vif(fit1))
   Outlets         RA         PI         AA Supervisors
 14.024686   7.271658   6.632887   1.991664   2.172542
>
> # remove Outlets
> fit2<-lm(Sales~RA+PI+AA+Supervisors,AS) #lm(Sales ~., data = AS[,c(-1,-3)])
> summary(fit2)

Call:
lm(formula = Sales ~ RA + PI + AA + Supervisors, data = AS)

Residuals:
       1        2        3        4        5        6        7        8        9       10 
 0.62145 -0.51436  0.81355  0.47026 -0.51176 -0.01728 -1.63989 -1.21176  1.79490  0.19489 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -19.11463    4.40661  -4.338  0.00744 ** 
RA            1.62834    0.26502   6.144  0.00166 ** 
PI            0.40055    0.01732  23.121 2.82e-06 ***
AA            2.00941    0.78449   2.561  0.05056 .  
Supervisors  -0.01545    0.15337  -0.101  0.92365    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.357 on 5 degrees of freedom
Multiple R-squared:  0.9943,	Adjusted R-squared:  0.9897 
F-statistic: 216.2 on 4 and 5 DF,  p-value: 8.731e-06

> (v2<-vif(fit2))
         RA          PI          AA Supervisors 
   2.058339    1.275934    1.960168    1.781990 
```

Removing insignificant variables in the model

```
> # remove insignificant contributor
> fit3<-lm(Sales~RA+PI+AA,AS) #lm(Sales ~., data = AS[,c(-1,-3,-7)])
> summary(fit3)

Call:
lm(formula = Sales ~ RA + PI + AA, data = AS)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.72068 -0.47520  0.09437  0.57302  1.73634 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -18.92388    3.63627  -5.204 0.002007 ** 
RA            1.61294    0.19785   8.152 0.000183 ***
PI            0.40031    0.01569  25.517 2.39e-07 ***
AA            1.96365    0.58458   3.359 0.015247 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.24 on 6 degrees of freedom
Multiple R-squared:  0.9942,	Adjusted R-squared:  0.9914 
F-statistic: 345.3 on 3 and 6 DF,  p-value: 4.17e-07

> (v3<-vif(fit3))
      RA       PI       AA 
1.373877 1.253012 1.303517 
```

Model selection summary

```
### summarising steps to selecting model
Names<-c("Intercept","No. retail outlets","No. registered automobiles","Personal income",
        "Average age of autos","No. supervisors")
Model<-c(rep(1,6),rep(2,5),rep(3,4))

# adjusted r squared output
r2_per1<-paste0(round(summary(fit1)$adj.r.squared*100,2),"%")
r2_per2<-paste0(round(summary(fit2)$adj.r.squared*100,2),"%")
r2_per3<-paste0(round(summary(fit3)$adj.r.squared*100,2),"%")
rsquared<-c(r2_per1,rep("",5),r2_per2,rep("",4),r2_per3,rep("",3))

Variables<-c(Names,Names[-2],Names[c(-2,-6)])

# regression coefficient
Coeff<-c(summary(fit1)$coefficients[,1],summary(fit2)$coefficients[,1],summary(fit3)$coefficients[,1])
Coeff<-round(Coeff,4)

# p-values
pvalues<-c("",ifelse(summary(fit1)$coefficients[-1,4]<0.0005,"<0.0005",
                     round(summary(fit1)$coefficients[-1,4],4)),
           "",ifelse(summary(fit2)$coefficients[-1,4]<0.0005,"<0.0005",
                      round(summary(fit2)$coefficients[-1,4],4)),
           "",ifelse(summary(fit3)$coefficients[-1,4]<0.0005,"<0.0005",
                      round(summary(fit3)$coefficients[-1,4],4)))
VIF<-c("",round(v1,4),"",round(v2,4),"",round(v3,4))

(Frame<-data.frame(Model,rsquared,Variables, Coeff,pvalues,VIF))
```

Model selection summary (output)

```
> (Frame<-data.frame(Model,rsquared,Variables, Coeff,pvalues,VIF))
   Model rsquared                Variables     Coeff  pvalues     VIF
1      1   98.72%                Intercept  -19.6715
2      1                  No. retail outlets  -0.0006   0.8234 14.0247
3      1            No. registered automobiles   1.7399   0.0346  7.2717
4      1                     Personal income   0.4099    7e-04  6.6329
5      1                Average age of autos   2.0357   0.0812  1.9917
6      1                     No. supervisors  -0.0344   0.8635  2.1725
7      2   98.97%                Intercept  -19.1146
8      2            No. registered automobiles   1.6283   0.0017  2.0583
9      2                     Personal income   0.4005  <0.0005  1.2759
10     2                Average age of autos   2.0094   0.0506  1.9602
11     2                     No. supervisors  -0.0155   0.9236   1.782
12     3   99.14%                Intercept  -18.9239
13     3            No. registered automobiles   1.6129  <0.0005  1.3739
14     3                     Personal income   0.4003  <0.0005   1.253
15     3                Average age of autos   1.9637   0.0152  1.3035
```

Standardised regression coefficients

```
> # standardised regression coefficients to compare coefficients
> library(QuantPsyc)
> (src1<-lm.beta(fit1)) # model 01
    Outlets          RA          PI          AA Supervisors
-0.03358557  0.31935538  0.90625417  0.12318168 -0.01016556
> src2<-lm.beta(fit2) # model 02
> src3<-lm.beta(fit3) # model 03
> # combine measurements into one vector/variables
> Standardised<-c("",round(src1,4),"",round(src2,4),"",round(src3,4))
> Frame<-mutate(Frame,Standardised) #append to model selection dataframe
> #### re-order variables
> (Frame<-Frame[,c(1:4,7,5:6)])
   Model rsquared                 Variables    Coeff Standardised pvalues     VIF
1      1   98.72%                 Intercept -19.6715
2      1               No. retail outlets  -0.0006      -0.0336  0.8234 14.0247
3      1          No. registered automobiles   1.7399       0.3194  0.0346  7.2717
4      1                   Personal income   0.4099       0.9063   7e-04  6.6329
5      1               Average age of autos   2.0357       0.1232  0.0812  1.9917
6      1                   No. supervisors  -0.0344      -0.0102  0.8635  2.1725
7      2   98.97%                 Intercept -19.1146
8      2          No. registered automobiles   1.6283       0.2989  0.0017  2.0583
9      2                   Personal income   0.4005       0.8855 <0.0005  1.2759
10     2               Average age of autos   2.0094       0.1216  0.0506  1.9602
11     2                   No. supervisors  -0.0155      -0.0046  0.9236   1.782
12     3   99.14%                 Intercept -18.9239
13     3          No. registered automobiles   1.6129       0.2961 <0.0005  1.3739
14     3                   Personal income   0.4003        0.885 <0.0005   1.253
15     3               Average age of autos   1.9637       0.1188  0.0152  1.3035
```

Stepwise regression

```
> ################################## stepwise regression
> library(MASS)
> # uses fit1 - i.e., the full model
> fit4 <- stepAIC(fit1, direction = "both", trace = F) #choose the best model by AIC
> summary(fit4)

Call:
lm(formula = Sales ~ RA + PI + AA, data = AS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72068 -0.47520  0.09437  0.57302  1.73634

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.92388    3.63627  -5.204 0.002007 **
RA            1.61294    0.19785   8.152 0.000183 ***
PI            0.40031    0.01569  25.517 2.39e-07 ***
AA            1.96365    0.58458   3.359 0.015247 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.24 on 6 degrees of freedom
Multiple R-squared:  0.9942,    Adjusted R-squared:  0.9914
F-statistic: 345.3 on 3 and 6 DF,  p-value: 4.17e-07
```

Final model summary

```
# sample output of overall result
Labels<-c("Overall","Intercept",
             "No. registed automobiles", "Personal income","Average age of automobiles")
(output<-summary(fit3))
################################## Overall
############## adjusted r squared
(r2<-output$adj.r.squared)
(r2_per<-paste0(round(r2*100,2),"%"))
(adj_r<-c(r2_per,"","","",""))
################################## regression coefficients
(Coeff<-c("",round(coefficients(fit3),4)))
################################## 95% CI (lower)
(Lower<-c("","",round(confint(fit3, level=0.95)[-1,1],4)))
################################## 95% CI (upper)
(Upper<-c("","",round(confint(fit3, level=0.95)[-1,2],4)))

################################## standardised regression coefficients
(Stand_Coeff<-c("","",round(src3,4)))
################################# p-values
############## overall p-value
lmp <- function (modelobject) {
  if (class(modelobject) != "lm") stop("Not an object of class 'lm' ")
  f <- summary(modelobject)$fstatistic
  p <- pf(f[1],f[2],f[3],lower.tail=F)
  attributes(p) <- NULL
  return(p)
}
(p<-ifelse(lmp(fit3)<0.0005,"<0.0005",round(lmp(fit3),4)))
############## individual p-values
(pvalues<-c(p,"",
      ifelse(output$coefficients[-1,4]<0.0005,"<0.0005",round(output$coefficients[-1,4],4))))
################################## dataframe
(df<-data.frame(Labels,adj_r,Coeff,Lower,Upper,Stand_Coeff,pvalues))
```

Final model summary (output)

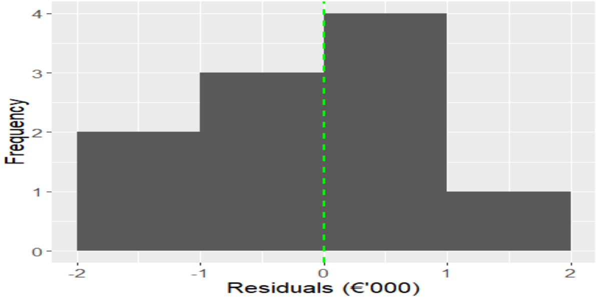
```
> ################################## dataframe
> (df<-data.frame(Labels,adj_r,Coeff,Lower,Upper,Stand_Coeff,pvalues))
                             Labels  adj_r    Coeff  Lower  Upper Stand_Coeff pvalues
                            Overall 99.14%                                     <0.0005
(Intercept)               Intercept        -18.9239
RA            No. registed automobiles       1.6129 1.1288 2.0971      0.2961 <0.0005
PI                  Personal income         0.4003 0.3619 0.4387       0.885 <0.0005
AA         Average age of automobiles        1.9637 0.5332 3.3941      0.1188  0.0152
```

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

```
# RESIDUALS
#normal
range(residuals(fit3))
(mn<-ggplot(fortify(fit3), aes(x = .resid)) + geom_histogram(breaks=seq(-2,2,1)))
(final_mn<-mn+labs(x="Residuals (€'000)", y="Frequency")+
    coord_cartesian(xlim=c(-2,2))+scale_x_continuous(breaks=seq(-2,2,1))+
    geom_vline(xintercept=0, linetype="dashed", color = "green", size=1)+
    theme(text = element_text(size=15)))
```



Residuals: Test of Normality/Differences

```
> shapiro.test(residuals(fit2)) #normality of residuals

        Shapiro-Wilk normality test

data:  residuals(fit2)
W = 0.98299, p-value = 0.9792

> t.test(residuals(fit2),mu=0)

        One Sample t-test

data:  residuals(fit2)
t = -1.1271e-16, df = 9, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.7237808  0.7237808
sample estimates:
    mean of x
-3.606327e-17
```

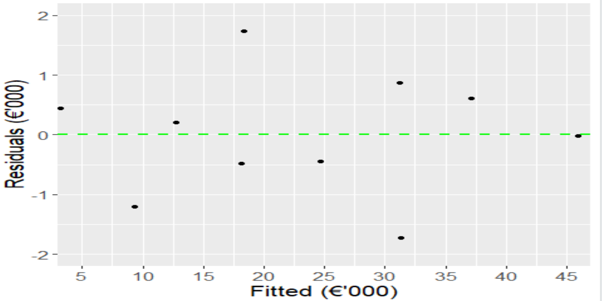
Residuals: Random

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
(mr<-ggplot(fortify(fit3), aes(x = .fitted, y = .resid)) + geom_point())
(final_mr<-mr+labs(x="Fitted (€'000)", y="Residuals (€'000)")+
    coord_cartesian(xlim=c(5,45),ylim=c(-2,2))+scale_x_continuous(breaks=seq(5,45,5))+
    scale_y_continuous(breaks=seq(-2,2,1))+
    geom_hline(yintercept=0, linetype="dashed", color = "green", size=1)+
    theme(text = element_text(size=15)))
```



Residuals: Durbin Watson test

```
> library(lmtest)
> dwtest(fit3) # autocorrelation of residuals

        Durbin-Watson test

data:  fit3
DW = 1.9772, p-value = 0.3234
alternative hypothesis: true autocorrelation is greater than 0
```

Guidelines for Multiple Regression

- A **Linear Relationship** between the outcome variable and the independent variables.
- **Multivariate Normality** - Multiple regression assumes that the variables are normally distributed.
- No **Multi-collinearity** - This assumption assumes that the independent variables are not highly correlated with each other.
- **Homoscedasticity** - This assumption requires that the variance of error terms are similar across the independent variables (i.e., the residuals are random).

Exercise 02: Multiple regression

A Department of Education analyst is interested in investigating the pay structure of secondary school teachers. He believes there are 3 factors that affect the salaries of teachers: years of experience; a rating of teaching effectiveness given by a department inspector; and the numbers of different subjects taught. A sample of 20 teachers produced the results given in the *Exercise02*.xlsx data set.

Using the steps outlined in the above slides to determine what measurements explain the variation in salaries of teachers.