Introduction to R_Slide 2

Dr. Ayat Almomani

Department of Statistics

Yarmouk University

12/16/2024

Import Data

- Read data from text file

```
Data_Revers=read.table("D_Rivers.txt", sep = "\t",
                       header = TRUE)
```

- Read data from Excel csv file

```
Data_Diabetes=read.csv(file="D_Diabetes.csv",sep=",",
                       header = TRUE)
```

- Read data from RData type

```
load("D_Diabetes.RData")
```

Data cleaning


```
attach(Data_Diabetes)
dim(Data_Diabetes)
```

```
## [1] 768  10
```

```
names(Data_Diabetes)
```

```
##  [1] "X"                        "Pregnancies"
##  [3] "Glucose"                  "BloodPressure"
##  [5] "SkinThickness"            "Insulin"
##  [7] "BMI"                      "DiabetesPedigreeF
##  [9] "Age"                      "Outcome"
```

Duplicates


```
sum(duplicated(Data_Diabetes))
```

```
## [1] 0
```

Missing Data


```
sum(is.na(Data_Diabetes))
```

```
## [1] 652
```

```
colSums(is.na(Data_Diabetes))
```

```
##                             X                   Pregnancies
##                             0                             0
##                 BloodPressure                 SkinThickness
##                            35                           227
##                           BMI DiabetesPedigreeFunction
##                            11                             0
##                       Outcome
##                             0
```

Drop missing data

- Drops rows missing values

```
dim(Data_Diabetes)
```

```
## [1] 768  10
```

```
D2=na.omit(Data_Diabetes)
dim(D2)
```

```
## [1] 392  10
```

Drop missing data

- Drops rows missing values for specific columns

```
library(tidyverse)
dim(Data_Diabetes)
```

```
## [1] 768  10
```

```
D3=drop_na(Data_Diabetes,SkinThickness,BMI)
dim(D3)
```

```
## [1] 539  10
```

Data Discriptive


```
str(Data_Revers)
```

```
## 'data.frame':    20 obs. of  6 variables:
##  $ River   : chr  "Olean" "Cassadaga" "Oatka" "Neversin
##  $ Agr     : int  26 29 54 2 3 19 16 40 28 26 ...
##  $ Forest  : int  63 57 26 84 27 61 60 43 62 60 ...
##  $ Rsdntial: num  1.2 0.7 1.8 1.9 29.4 3.4 5.6 1.3 1.1
##  $ ComIndl : num  0.29 0.09 0.58 1.98 3.11 0.56 1.11 0.
##  $ Nitrogen: num  1.1 1.01 1.9 1 1.99 1.42 2.04 1.65 1.
```

Data Discriptive


```
summary(Data_Revers)
```

```
##     River                Agr           Forest          Rsdntial      
##  Length:20          Min.   : 2.0   Min.   :26.00   Min.   : 0.400  
##  Class :character   1st Qu.: 5.5   1st Qu.:55.25   1st Qu.: 0.700  
##  Mode  :character   Median :20.0   Median :61.50   Median : 0.900  
##                     Mean   :19.4   Mean   :62.85   Mean   : 2.715  
##                     3rd Qu.:26.5   3rd Qu.:76.50   3rd Qu.: 1.425  
##                     Max.   :54.0   Max.   :89.00   Max.   :29.400  
##     ComIndl          Nitrogen      
##  Min.   :0.0900   Min.   :0.660  
##  1st Qu.:0.1500   1st Qu.:0.800  
##  Median :0.2250   Median :1.010  
##  Mean   :0.5155   Mean   :1.157  
##  3rd Qu.:0.4025   3rd Qu.:1.353  
##  Max.   :3.1100   Max.   :2.040
```

Data Discriptive

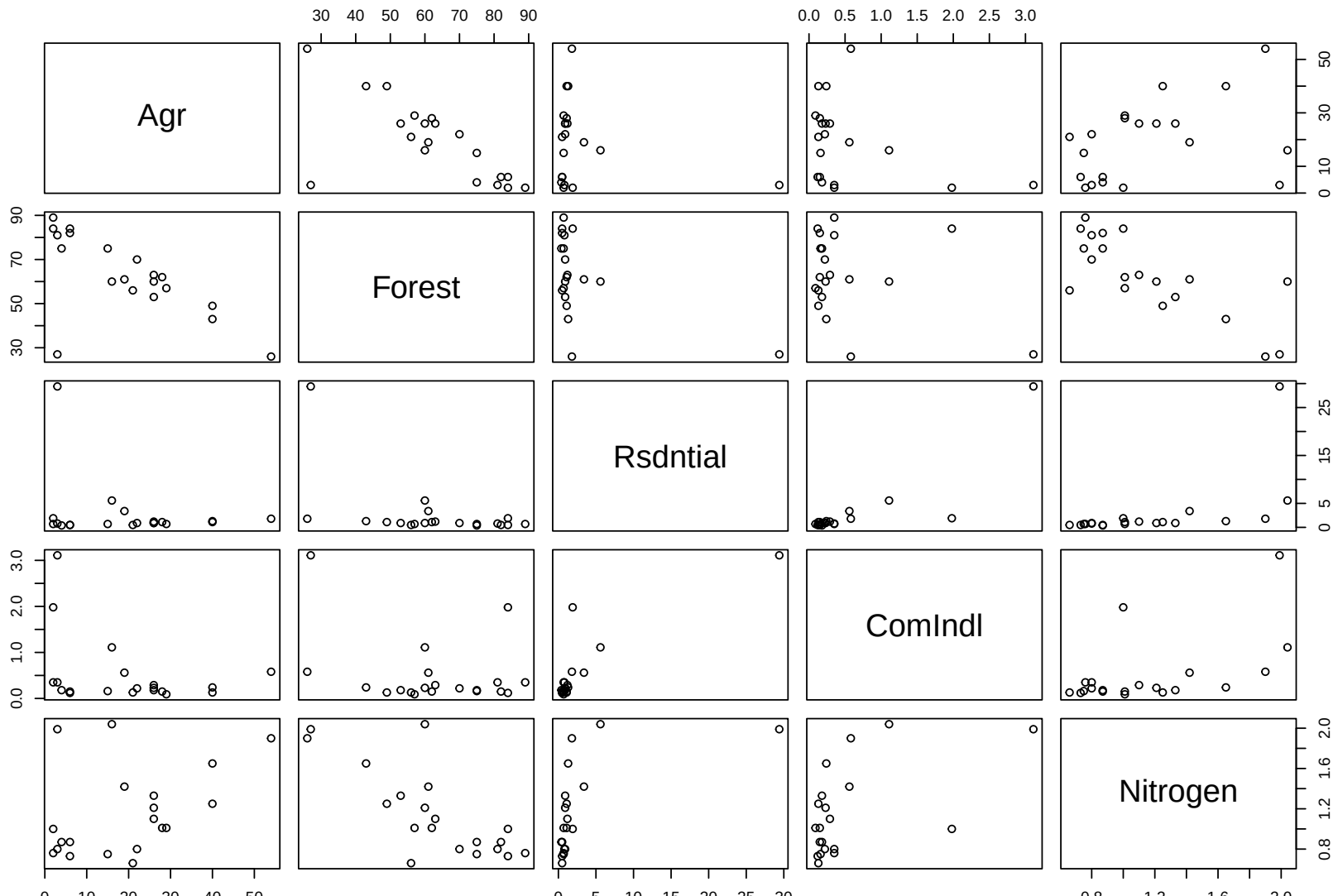

```
library(psych)
describe(Data_Revers[,-1])
```

```
##           vars  n  mean    sd median trimmed   mad   min   max range  skew
## Agr          1 20 19.40 14.73  20.00   18.12 17.05  2.00 54.00 52.00  0.52
## Forest       2 20 62.85 17.84  61.50   64.44 19.27 26.00 89.00 63.00 -0.48
## Rsdntial     3 20  2.71  6.40   0.90    1.15  0.52  0.40 29.40 29.00  3.61
## ComIndl      4 20  0.52  0.76   0.22    0.31  0.14  0.09  3.11  3.02  2.35
## Nitrogen     5 20  1.16  0.44   1.01    1.11  0.36  0.66  2.04  1.38  0.80
##           kurtosis   se
## Agr          -0.60 3.29
## Forest       -0.60 3.99
## Rsdntial     12.14 1.43
## ComIndl       4.72 0.17
## Nitrogen     -0.73 0.10
```

Data Discriptive

Dr. Ayat Almomani

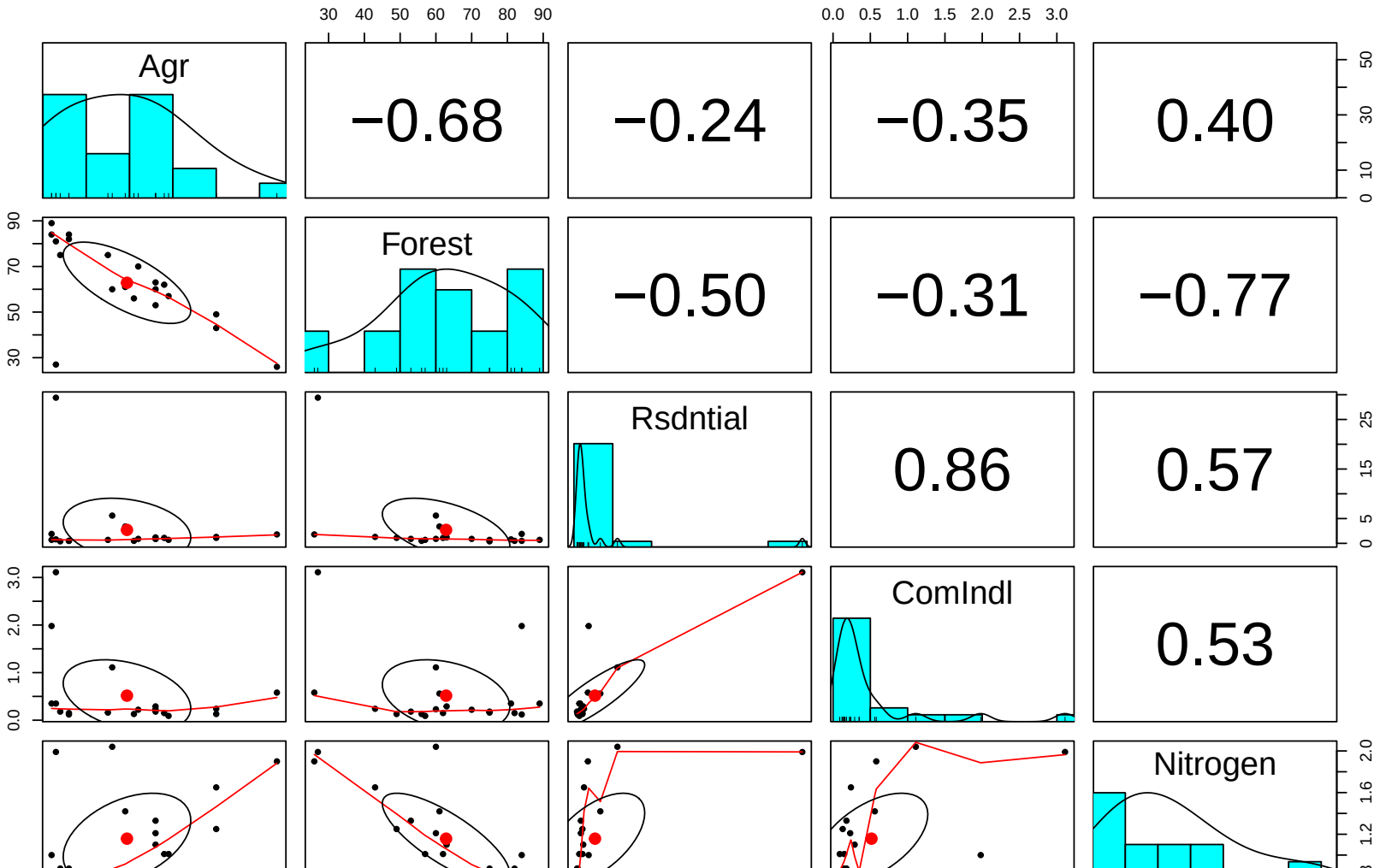
```
pairs(Data_Revers[,2:6])
```



Data Discriptive

Dr. Ayat Almomani

```
library(psych)
pairs.panels(Data_Revers[,-1],col="red")
```



Data Discriptive


```
z=round(cor(Data_Revers[,-1]),2)
z
```

```
##                Agr Forest Rsdntial ComIndl Nitrogen
## Agr           1.00  -0.68    -0.24   -0.35     0.40
## Forest       -0.68   1.00    -0.50   -0.31    -0.77
## Rsdntial     -0.24  -0.50     1.00    0.86     0.57
## ComIndl      -0.35  -0.31     0.86    1.00     0.53
## Nitrogen      0.40  -0.77     0.57    0.53     1.00
```

```
attach(Data_Revers)
cor(Agr,Nitrogen)
```

```
## [1] 0.400951
```

Correlation test


```
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         conf.level = 0.95,  ...)
```

Correlation test

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

```
cor.test(Agr,Nitrogen)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Agr and Nitrogen
## t = 1.8569, df = 18, p-value = 0.07977
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05053647  0.71636731
## sample estimates:
##      cor
## 0.400951
```

Export Data

- Export **txt** data type

```
x=round(describe(Data_Revers[,-1]),2)
write.table(x, file="Data_Descriptive.txt", sep="\t",
            row.names=FALSE)
```

- Export **csv** data type

```
write.csv(x, file="Data_Descriptive.csv")
```

- Export txt data type

```
save(Data_Revers, x, file = "Data2.RData")
```

The Regression model

- The Model equation

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_1 x_{pi} + \epsilon_i; \ i = 1, \ldots, n$$

- Using R, the function **lm()** is used to fit the model.
- The function **summary()** is used to summarize the model.
- The function **plot()** is used to check the assumptions.

Fitting the Regression model


```
Rivers_LM=lm(Nitrogen ~ Agr+Forest+Rsdntial+
                  ComIndl,data = Data_Revers)
```

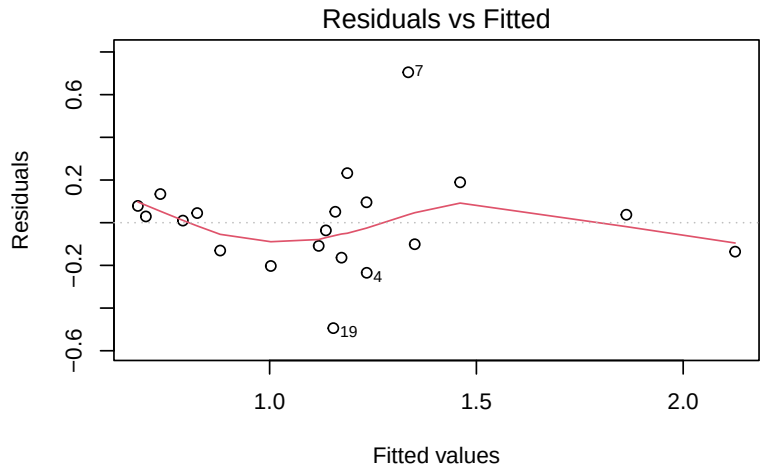
```
summary(Rivers_LM)
```

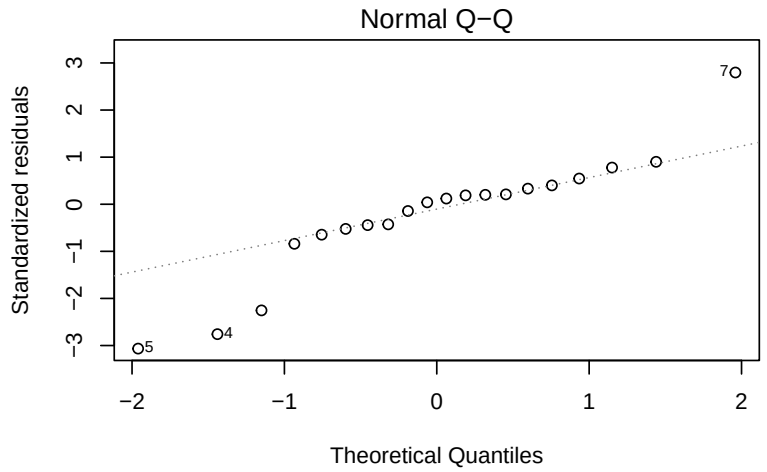
```
##
## Call:
## lm(formula = Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = Data_Revers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49404 -0.13180  0.01951  0.08287  0.70480
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.722214   1.234082   1.396   0.1832
## Agr          0.005809   0.015034   0.386   0.7046
## Forest      -0.012968   0.013931  -0.931   0.3667
## Rsdntial    -0.007227   0.033830  -0.214   0.8337
## ComIndl      0.305028   0.163817   1.862   0.0823 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2649 on 15 degrees of freedom
## Multiple R-squared:  0.7094, Adjusted R-squared:  0.6319
## F-statistic: 9.154 on 4 and 15 DF,  p-value: 0.0005963
```

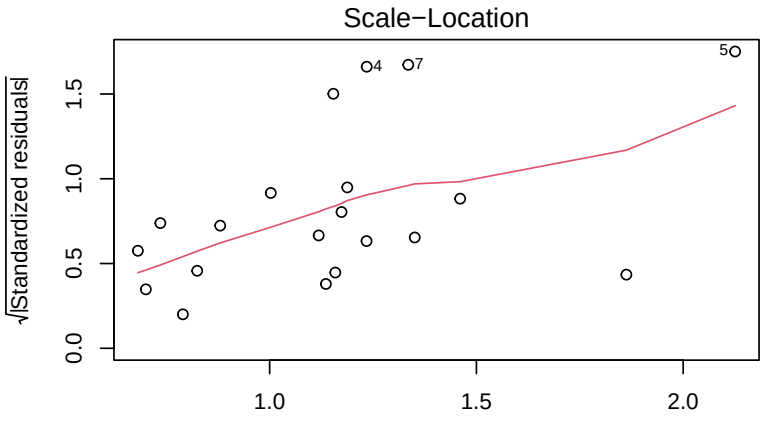
Check the Assumptions

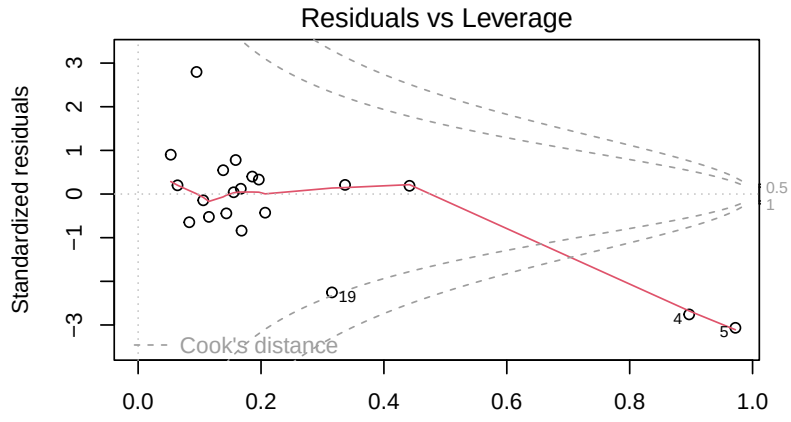
Dr. Ayat Almomani

```
par(mfrow = c(2, 2))
plot(Rivers_LM)
```





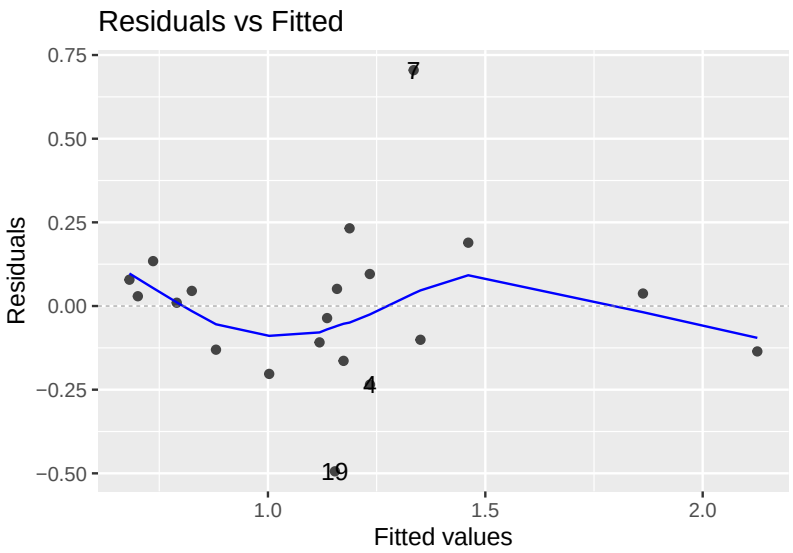


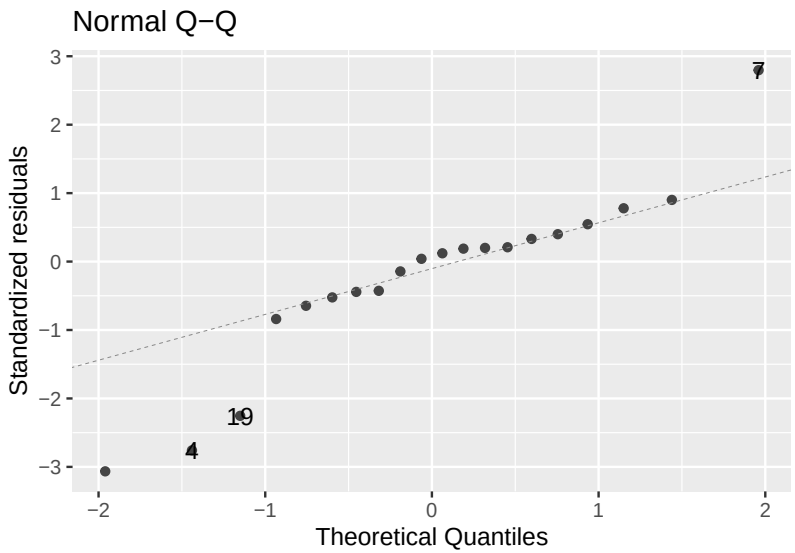


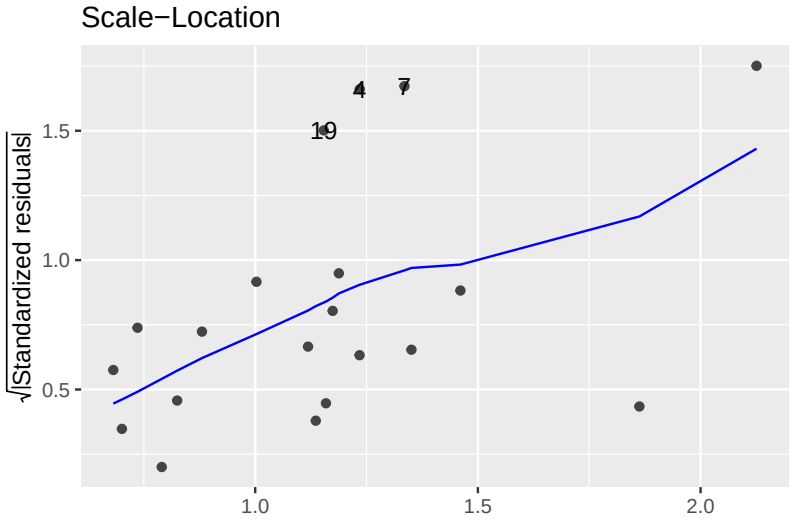
Check the Assumptions

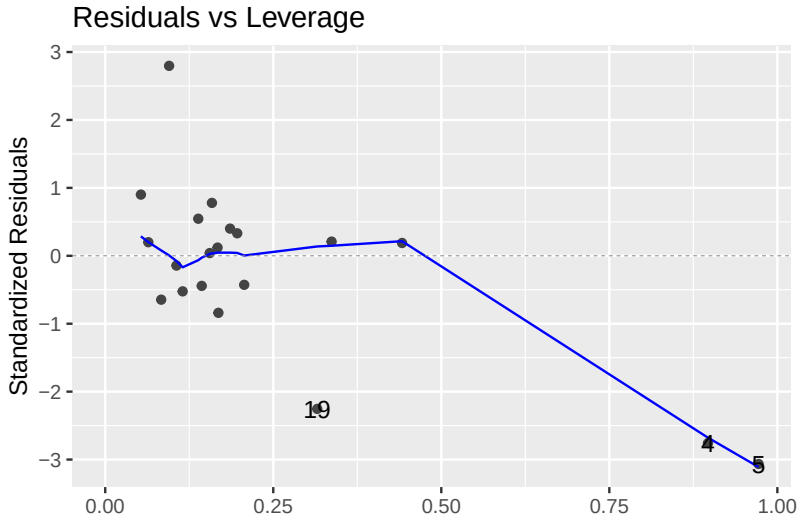
Dr. Ayat Almomani

```
library(ggfortify)
autoplot(Rivers_LM)
```









Formulas

Dr. Ayat Almomani

Model	Interpretation
`y ~ 1`	Just the intercept
`y ~ a`	One main effect
`y ~ -1+a`	No intercept
`y ~ a+b`	Two main effects
`y ~ a+b+c+a:b`	Three main effects and an interaction between `a` and `b`
`y ~ a*b`	All main effects and interactions (same as `a+b+a:b`)
`y ~ factor(a)`	Create dummy variables for `a` (if not already a factor)
`y ~ (a+b+c)^2`	All main effects and second-order interactions
`y ~ I(a^2)`	Transform `a` to `a^2`
`log(y) ~ a`	Log transform `y`
`y ~ a/b/c`	Factor `c` nested within factor `b` within factor `a`
`y ~ .`	Main effect for each column in the dataframe

Qualitative Independent Variables

- Data on salaries of employees in IT (many years ago) based on their years of experience, their education level and whether or not they are management.
- Outcome: S, salaries for IT staff in a corporation.
- Predictors:
  - X, experience (years)
  - E, education (1=Bachelor's, 2=Master's, 3=Ph.D)
  - M, management (1=management, 0=not management)

Data


```
salary_Data <- read.table("D_salary.table", header=T)
salary_Data$E <- factor(salary_Data$E)
salary_Data$M <- factor(salary_Data$M)
summary(salary_Data)
```

```
##        S                X         E       M
##  Min.   :10535   Min.   : 1.0   1:14   0:26
##  1st Qu.:13321   1st Qu.: 3.0   2:19   1:20
##  Median :16436   Median : 6.0   3:13
##  Mean   :17270   Mean   : 7.5
##  3rd Qu.:20720   3rd Qu.:11.0
##  Max.   :27837   Max.   :20.0
```

Factor variables


```
head(salary_Data$E)
```

```
## [1] 1 3 3 2 3 2
## Levels: 1 2 3
```

```
head(salary_Data$M)
```

```
## [1] 1 0 1 0 0 1
## Levels: 0 1
```

Define qualitative variables

- We first define qualitative variables related to E:
  - $E_{i2} = 1$, if Master's; 0 otherwise
  - $E_{i3} = 1$, if PhD; 0 otherwise
- In fact, R is doing it automatically for us as you can see from the output from regression:

$$S_i = \beta_0 + \beta_1 X_i + \beta_2 E_{i2} + \beta_3 E_{i3} + \beta_4 M_i + \epsilon_i$$

Regression Model


```
Salary_lm <- lm(S ~ E + M + X, salary_Data)
summary(Salary_lm )
```

```
##
## Call:
## lm(formula = S ~ E + M + X, data = salary_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1884.60  -653.60    22.23   844.85  1716.47
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8035.60     386.69  20.781  < 2e-16 ***
## E2           3144.04     361.97   8.686 7.73e-11 ***
## E3           2996.21     411.75   7.277 6.72e-09 ***
## M1           6883.53     313.92  21.928  < 2e-16 ***
## X             546.18      30.52  17.896  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1027 on 41 degrees of freedom
## Multiple R-squared:  0.9568, Adjusted R-squared:  0.9525
## F-statistic: 226.8 on 4 and 41 DF,  p-value: < 2.2e-16
```

Regression Equation

- For Bachelor's and Management track, the regression equation is $E(S) = (\beta_0 + \beta_4) + \beta_1 X$
- For Bachelor's and non-Management track, the regression equation is $E(S) = (\beta_0) + \beta_1 X$
- For Master's and Management track, the regression equation is $E(S) = (\beta_0 + \beta_2 + \beta_4) + \beta_1 X$
- For Master's and non-Management track, the regression equation is $E(S) = (\beta_0 + \beta_2) + \beta_1 X$
- For PhD and Management track, the regression equation is $E(S) = (\beta_0 + \beta_3 + \beta_4) + \beta_1 X$
- For PhD and non-Management track, the regression equation is $E(S) = (\beta_0 + \beta_3) + \beta_1 X$

Design Matrix


```
head(model.matrix(Salary_lm))
```

```
##   (Intercept) E2 E3 M1 X
## 1           1  0  0  1 1
## 2           1  0  1  0 1
## 3           1  0  1  1 1
## 4           1  1  0  0 1
## 5           1  0  1  0 1
## 6           1  1  0  1 2
```

Interaction terms between X and E

$$S_i = \beta_0 + \beta_1 X_i + \beta_2 E_{i2} + \beta_3 E_{i3} + \beta_4 M_i + \beta_5 E_{i2} X_i + \beta_6 E_{i3} X_i + \epsilon_i$$

```
Salary_lm2<- lm(S~ E + M + X + X:E, salary_Data)
summary(Salary_lm2)
```

```
##
## Call:
## lm(formula = S ~ E + M + X + X:E, data = salary_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2013.04  -634.68   -16.71   615.66  2014.14
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7256.28     549.49  13.205 5.65e-16 ***
## E2           4172.50     674.97   6.182 2.90e-07 ***
## E3           3946.36     686.69   5.747 1.16e-06 ***
## M1           7102.45     333.44  21.300  < 2e-16 ***
## X             632.29      53.19  11.888 1.53e-14 ***
## E2:X         -125.51      69.86  -1.797   0.0801 .
## E3:X         -141.27      89.28  -1.582   0.1216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1005 on 39 degrees of freedom
## Multiple R-squared:  0.9606, Adjusted R-squared:  0.9546
```

Compare models


```
anova(Salary_lm, Salary_lm2)
```

```
## Analysis of Variance Table
##
## Model 1: S ~ E + M + X
## Model 2: S ~ E + M + X + X:E
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     41 43280719
## 2     39 39410680  2   3870040 1.9149  0.161
```

Design matrix


```
model.matrix(Salary_lm2)[10:20,]
```

```
##    (Intercept) E2 E3 M1 X E2:X E3:X
## 10           1  1  0  0 3    3    0
## 11           1  0  0  1 3    0    0
## 12           1  1  0  1 3    3    0
## 13           1  0  1  1 3    0    3
## 14           1  0  0  0 4    0    0
## 15           1  0  1  1 4    0    4
## 16           1  0  1  0 4    0    4
## 17           1  1  0  0 4    4    0
## 18           1  1  0  0 5    5    0
## 19           1  0  1  0 5    0    5
## 20           1  0  0  1 5    0    0
```

Interaction of E and M:

$$S_i = \beta_0 + \beta_1 X_i + \beta_2 E_{i2} + \beta_3 E_{i3} + \beta_4 M_i + \beta_5 E_{i2} M_i + \beta_6 E_{i3} M_i + \epsilon_i$$

```
Salary_lm3<-lm(S~ E + M + X + M:E, salary_Data)
summary(Salary_lm3)
```

```
##
## Call:
## lm(formula = S ~ E + M + X + M:E, data = salary_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -928.13  -46.21   24.33   65.88  204.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9472.685     80.344  117.90   <2e-16 ***
## E2          1381.671     77.319   17.87   <2e-16 ***
## E3          1730.748    105.334   16.43   <2e-16 ***
## M1          3981.377    101.175   39.35   <2e-16 ***
## X            496.987      5.566   89.28   <2e-16 ***
## E2:M1       4902.523    131.359   37.32   <2e-16 ***
## E3:M1       3066.035    149.330   20.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173.8 on 39 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9986
```

ANOVA


```
anova(Salary_lm, Salary_lm3)
```

```
## Analysis of Variance Table
##
## Model 1: S ~ E + M + X
## Model 2: S ~ E + M + X + M:E
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     41 43280719
## 2     39  1178168  2  42102552 696.84 < 2.2e-16 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '
```

Diagnostics

Dr. Ayat Almomani

```
par(mfrow=c(2,2))
plot(Salary_lm3)
```