UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

SMDE REPORT – SECOND ASSIGNMENT

Mohana Fathollahi
Jun 15, 2021

Steps to do this assignment:

1.Generate a dataset with ten factors, the first five factors F1, F2, F3, F4 and F5 have been generated from different distributions and the factors F6, F7, F8, F9 and F10 generated by linear combinations of the first five factors. R has been used for this part.

- f1 <- rnorm(2000, mean = 5 , sd = 3)
- f2 <- rexp(2000,rate = 20)
- f3 <- rnorm(2000, mean = 3 , sd = 1.5)
- f4 <- rnorm(2000, mean = 0 , sd = 1)
- f5 <- rexp(2000,rate = 10)
-
- f6 <- f5+f3
- f7 <- 2*f5+ 3*f1
- f8 <- f4+ 2.5*f2+11*f3
- f9 <- f3+ 6*f2
- f10 <- f5+2*f4

2. Define an answer variable that is composed by a function combining a subset of the ten factors plus a normal distribution (to add random noise). R has been used for this part.

Answer <- 3* f1 + 4*f7 +10* f5 + $\mathcal{N}(0, 1)$.

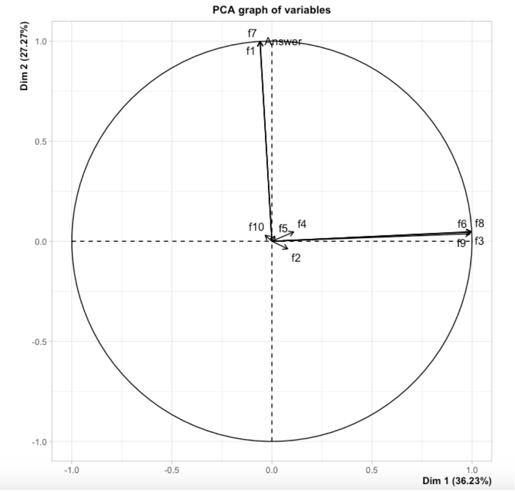
I will consider that the answer variable (called time to produce a product) represents the amount of time needed to do each step.

3. Proposing an expression to understand the relations between the data (model).

To analyse and understand the data generated in the answer variable a Principal Component Analysis was proposed. Based on the result of running PCA we could find that **72.955** percent of the variance is explained by the first three principal components.

```
         Eigenvalues

              Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9

Variance              3.985   3.000   1.040   1.024   0.997   0.953   0.000   0.000   0.000

% of var.            36.229  27.270   9.457   9.312   9.067   8.663   0.003   0.000   0.000

Cumulative % of var.  36.229  63.499  72.955  82.267  91.335  99.997 100.000 100.000 100.000
```

To obtain a better understanding of the data PCA gives us this plot.



To find that each component consists of which factors, we should find the highest weight in each first tree component.

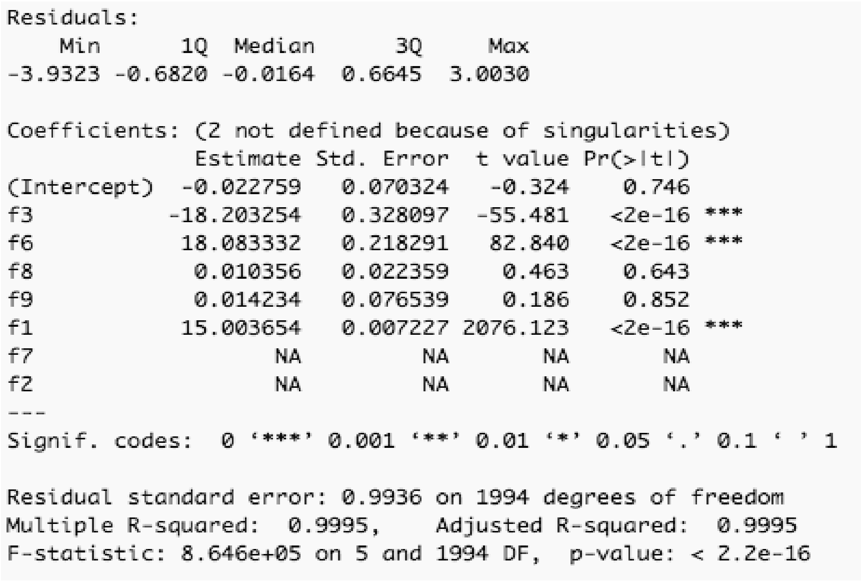
```
                          Variables (the 10 first)

               Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2

f1     | -0.060  0.090  0.004 |  0.998 33.171  0.995 |  0.002  0.000  0.000 |

f2     |  0.079  0.156  0.006 | -0.038  0.048  0.001 |  0.827 65.708  0.684 |

f3     |  0.995 24.865  0.991 |  0.045  0.069  0.002 | -0.049  0.233  0.002 |

f4     |  0.109  0.298  0.012 |  0.047  0.073  0.002 | -0.339 11.048  0.115 |

f5     |  0.015  0.005  0.000 |  0.026  0.022  0.001 |  0.422 17.132  0.178 |
f6    |  0.995 24.828  0.989 |  0.047  0.074  0.002 | -0.021  0.042  0.000 |
f7    | -0.060  0.089  0.004 |  0.998 33.198  0.996 |  0.011  0.013  0.000 |
f8    |  0.997 24.922  0.993 |  0.048  0.076  0.002 | -0.063  0.382  0.004 |
f9    |  0.991 24.630  0.982 |  0.037  0.046  0.001 |  0.108  1.120  0.012 |
f10   | -0.033  0.028  0.001 |  0.030  0.030  0.001 |  0.211  4.288  0.045 |
```

These components are configured as follows:

- First component: Factors 3, 6, 8 and 9.
- Second component: Factors 1, 7.
- Third component: Factors 2.

We will use all these factors to build a model.

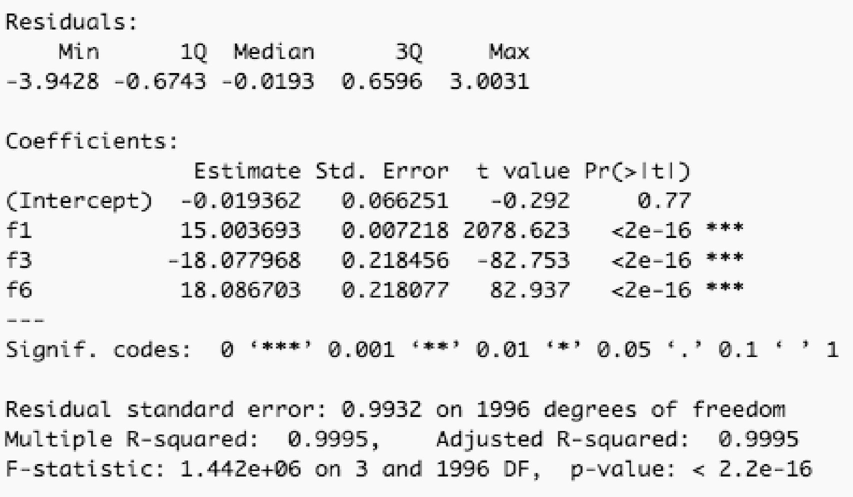
model_1 <-lm(Answer~f3+f6+f8+f9+f1+f7+f2)



```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.9323 -0.6820 -0.0164  0.6645  3.0030 

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  -0.022759   0.070324   -0.324    0.746    
f3          -18.203254   0.328097  -55.481   <2e-16 ***
f6           18.083332   0.218291   82.840   <2e-16 ***
f8            0.010356   0.022359    0.463    0.643    
f9            0.014234   0.076539    0.186    0.852    
f1           15.003654   0.007227 2076.123   <2e-16 ***
f7                  NA         NA       NA       NA    
f2                  NA         NA       NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9936 on 1994 degrees of freedom
Multiple R-squared:  0.9995,	Adjusted R-squared:  0.9995 
F-statistic: 8.646e+05 on 5 and 1994 DF,  p-value: < 2.2e-16
```

Based on the result of the first model, factors 7 and 4 do not present valid results, so we remove them from the model. Additionally, factor 8 and 9 do not seem to have too much influence in the presence of the other variables because its p-value is non significant. We should remove this factor too. Therefore our second model will be like this:

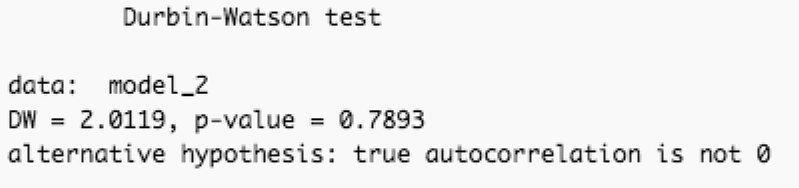
model_2 <-lm(Answer~f3+f6+f1)



```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.9428 -0.6743 -0.0193  0.6596  3.0031 

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  -0.019362   0.066251   -0.292     0.77    
f1           15.003693   0.007218 2078.623   <2e-16 ***
f3          -18.077968   0.218456  -82.753   <2e-16 ***
f6           18.086703   0.218077   82.937   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9932 on 1996 degrees of freedom
Multiple R-squared:  0.9995,	Adjusted R-squared:  0.9995 
F-statistic: 1.442e+06 on 3 and 1996 DF,  p-value: < 2.2e-16
```

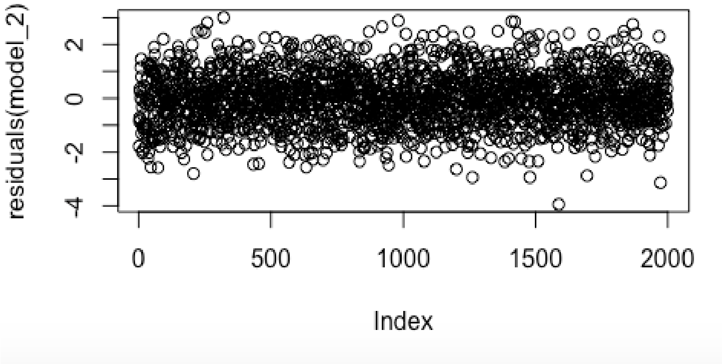
This time all variables are statistical significant and the adjusted R2 did not change. The next step is checking assumptions:

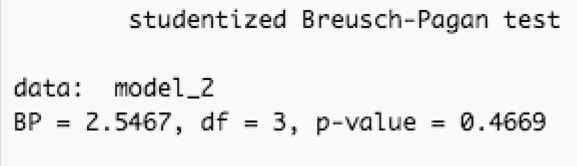
1. The p-value > 0.05. Therefore, we accept Independence of observations.



```
	Durbin-Watson test

data:  model_2
DW = 2.0119, p-value = 0.7893
alternative hypothesis: true autocorrelation is not 0
```

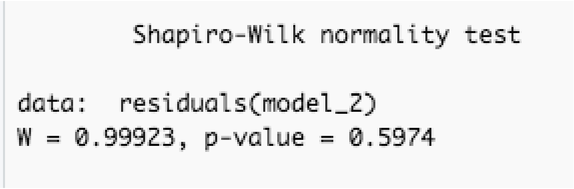
2. Homogeneity of variance, p-value is larger than 0.05 so we accept Homogeneity of variances.



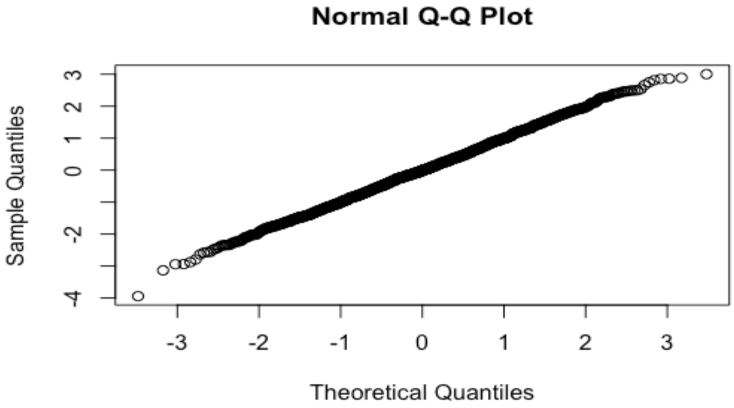


```
	studentized Breusch-Pagan test

data:  model_2
BP = 2.5467, df = 3, p-value = 0.4669
```

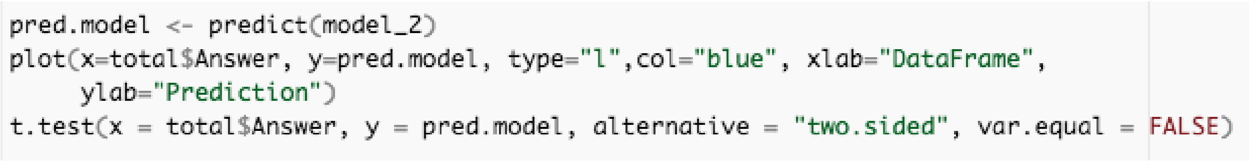
3. Normality test for these three factors, p_vlue for these factors are larger than 0.05 so we accept the null hypothesis.



```
	Shapiro-Wilk normality test

data:  residuals(model_2)
W = 0.99923, p-value = 0.5974
```

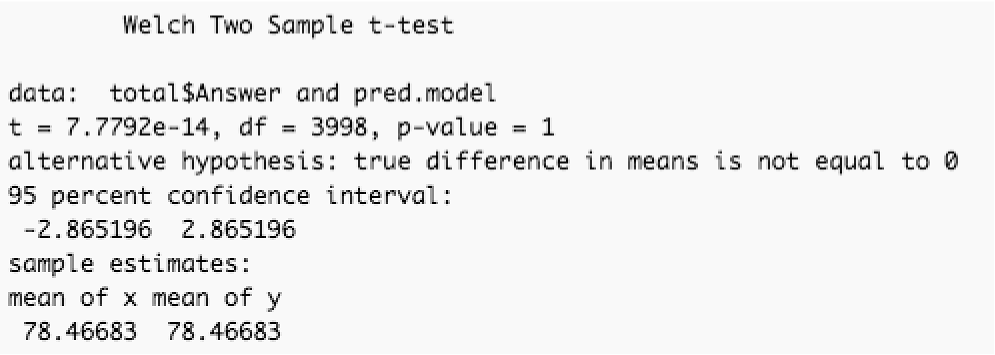


To validate the results of our model. Using the model to predict the data and compare them with the actual data using a t-test and ANOVA.



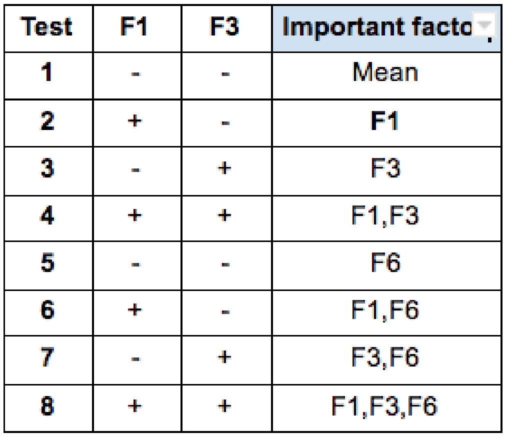
```
pred.model <- predict(model_2)
plot(x=total$Answer, y=pred.model, type="l",col="blue", xlab="DataFrame",
     ylab="Prediction")
t.test(x = total$Answer, y = pred.model, alternative = "two.sided", var.equal = FALSE)
```

Based on the result of this, we found that p-value=1.Therefore, there is not a statistically significant difference in the means of the predicted and real data. All parts have been done in R.



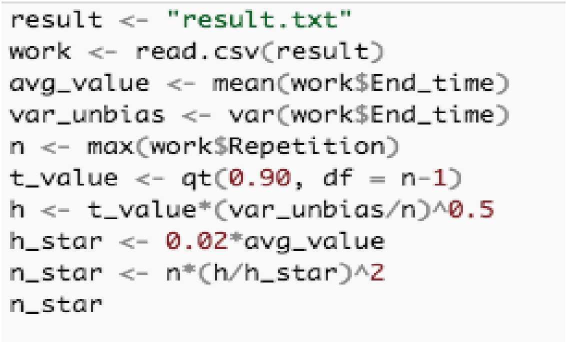
```
	Welch Two Sample t-test

data:  total$Answer and pred.model
t = 7.7792e-14, df = 3998, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.865196  2.865196
sample estimates:
mean of x mean of y 
 78.46683  78.46683 
```

4.In this step we should generate new data, GPSS has been used to generate data in different scenarios, like the table below. We assumed that each of the factors in the selected model would represent the operation of a machine in a production line. Additionally, normal distributions are added to each variable to generate noise in the results.



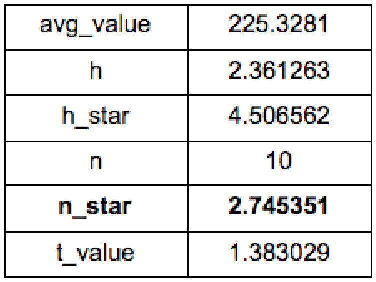
Test	F1	F3	Important facto
1	-	-	Mean
2	+	-	**F1**
3	-	+	F3
4	+	+	F1,F3
5	-	-	F6
6	+	-	F1,F6
7	-	+	F3,F6
8	+	+	F1,F3,F6

Next that we should find number of repetition, first we will consider n= 10 and by result of GPSS model we run this code in R:



```
result <- "result.txt"
work <- read.csv(result)
avg_value <- mean(work$End_time)
var_unbias <- var(work$End_time)
n <- max(work$Repetition)
t_value <- qt(0.90, df = n-1)
h <- t_value*(var_unbias/n)^0.5
h_star <- 0.02*avg_value
n_star <- n*(h/h_star)^2
n_star
```

Based on the result of this code,table below, we can find that n_star = 3 so we need 3 repetition for each scenario.



avg_value	225.3281
h	2.361263
h_star	4.506562
n	10
n_star	**2.745351**
t_value	1.383029

5.Define a DoE to explore with what parametrization of the factors the answer variable obtains the best results, analysing interactions.

In this part Yates algorithm has been used. The most effective factor has been found through this method. All processes have been done in excel and attached to this report. Based on the results, table below, we can conclude that the main effect is obtained by changing the processing time of the factor1. But in this case we need to reduce time so minous values are more important. F3 has larger value and after that interaction of F3 and F6 has larger value. Therefore these factors should be considered as important ones.

Effects	***Important factor***
280.0751001	Mean
10.6115458	**F1**
-6.346635392	**F3**
0.6672801238	F1,F3
-0.8919125092	F6
-3.439980082	F1,F6
-4.035239524	**F3,F6**
2.677593223	F1,F3,F6