Ömer Yiğit – r0767950
Prof. Thomas Neyens (2019 – 2020)

REGRESSION ANALYSIS – EXAM PROJECT

1) Exploratory analyses were conducted on the training data. First, there was no missing data. According to visual inspection, there were no significant irregularities regarding the outliers and skewness. Examining the standardized data, there were only two points (1 and 355) that have values outside ±3σ limits. Since there is no information about the sampling process, they are kept in the dataset. The next step was to obtain the correlation matrix, which is given in *Table 1*. According to the values, two highest correlations occur between (SWI, SWF) pair and (temperature, duration) pair. However, they are not influential, in terms of multicollinearity.

	SWI	**SWF**	**temperature**	**size**	**management**	**duration**
SWI	1.0000	0.6803	0.3768	0.0924	0.3814	0.1992
SWF	0.6803	1.0000	-0.0497	0.0445	0.1636	-0.0854
temperature	0.3768	-0.0497	1.0000	0.0330	0.0967	0.7708
size	0.0924	0.0445	0.0330	1.0000	0.0652	-0.0251
management	0.3814	0.1636	0.0967	0.0652	1.0000	0.0386
duration	0.1992	-0.0854	0.7708	-0.0251	0.0386	1.0000

Table 1: Correlation matrix

2) A linear first-order regression model is fitted on the data. SWI is the response variable; SWF, temperature, size and management are the predictor variables.

$$\widehat{SWI} = -0.6397 + 0.9076 \times SWF + 0.0546 \times temp. + 0.0014 \times size + 0.0645 \times manag.$$

(a) While the size variable is not significant, the overall model fits the data well and it explains around 68% of the variation. This is quite adequate, yet it can be improved.

(b) Gauss-Markov conditions are checked (except independence) and it is found that there is heteroscedasticity in the residuals of the fitted model. The other conditions are satisfied.

(c) Multicollinearity is checked. VIF (variation inflation factor) values for all four variables in the model are around 1, which are smaller than 10, so there is no multicollinearity detected. Also, the result coincides according to the condition numbers, none of which is above 30.

(d) Studentized residuals are checked for influential outliers. There is only one observation outside the limits. However, according to Cook's distance, there are no outliers. Also, according to DFFITS and DFBETAS calculations, which are more powerful, there are 15 influential points. In general, using *influence.measures* function, 3 influential points (1, 27 and 43) are observed.
Another approach is to use a robust diagnostic plot. A robust regression model is fitted with *ltsReg* function (with using the same variables of the model given above). The plot is given in *Figure 1*.

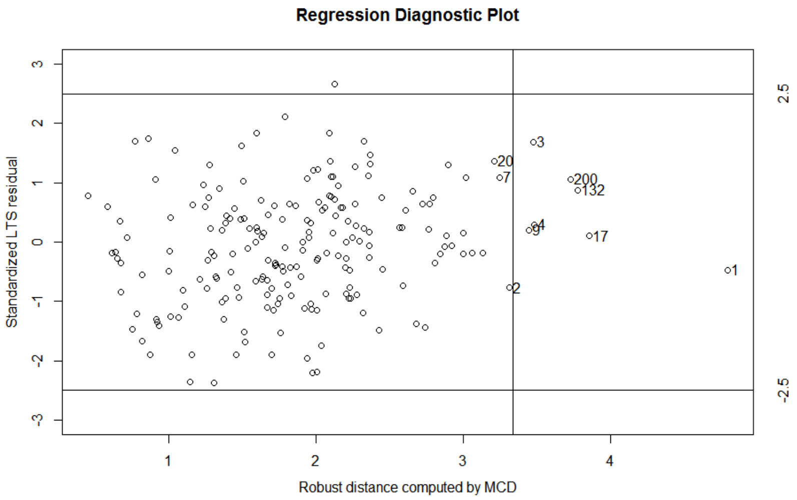

Figure 1: Robust diagnostic plot for the model

According to the plot, there are seven good leverage points and only one vertical outlier (355). This coincides with the studentized residuals of the non-robust linear model. Hence, in conclusion, it is decided that there is only one significant outlier point. However, since there is no information about the sampling process, this point is not removed for further analyses.

3) Heteroscedasticity of the residuals is the problem to be remedied. As the first step, stepwise methods are tried. In all three methods (backward elimination, forward selection and stepwise regression), the final models are the same – only SWF, temperature and management variables are used. However, the problem is not remedied with this model. Residual vs. Fitted plot and Scale-Location plots suggest that new variable(s) are needed. Therefore, transformations are tried.

- *Response variable transformations*: Logarithmic transformation did not solve the problem. Square-root and box-cox transformations (with λ = 0.645) solved it.
- *Explanatory variable transformations*: Different combinations of logarithmic and square root transformations are applied to the explanatory variables. However, there were no significant improvement nor solution to heteroscedasticity problem. Then, different combinations of polynomial – squared and cubic – transformations are applied. The only applicable transformation that improves the model and satisfy Gauss-Markov conditions is square transformation. The new variable is selected to be SWF^2.

Even though SWF is not significant, it is kept in the models to prevent distortion in the refits. In total, there are three applicable transformations that have significant explanatory power and that satisfy the Gauss-Markov conditions. With their combinations, five new models are created. Additionally, weighted least square regression approach is tried on each model. Weights of the variables are inversely proportional to their variances and they are recalculated for each model, otherwise the Gauss-Markov conditions would not be satisfied. In total, 10 models are created from which the best model is to be selected according to certain criteria. A summary of the comparison procedure is given in *Table 2*. Names of the used models for the analyses are based on R code, which can be found in the Appendix.

	UNWEIGHTED				**WEIGHTED**			
Model	**Res. Std.**	**Adj. R^2**	**AIC**	**PRESS**	**Res. Std.**	**Adj. R^2**	**AIC**	**PRESS**
lm3	0.1416	0.6805	-208.4510	4.0881	1.2490	0.7064	-213.8931	4.0982
lm5	0.3096	0.6831	104.4839	19.5456	1.2470	0.7082	100.6240	19.5717
lm6	0.3845	0.7021	192.1589	30.2431	1.2460	0.7568	181.3899	30.1453
lm7	0.1397	0.6889	-212.7907	3.9893	1.2760	0.7633	-217.4489	3.9844
lm8	0.3038	0.6949	97.9020	18.8565	1.2660	0.7585	93.7920	18.8325

Table 2: Comparison of the suitable improved models

The most suitable models seem to be lm3 and lm7, because in both unweighted and weighted cases, they have the lowest values for AIC and PRESS. Even though weighted models explain more variability, unweighted models have lower variance of the residuals and are easier to interpret. Hence, in conclusion, lm3 and lm7 are selected as the improved model choices. The models are as follows:

$$lm3 : \widehat{\sqrt{SWI}} = \beta_0 + \beta_1 \times SWF + \beta_2 \times temperature + \beta_3 \times management$$

$$lm7 : \widehat{\sqrt{SWI}} = \beta_0 + \beta_1 \times SWF + \beta_2 \times SWF^2 + \beta_3 \times temperature + \beta_4 \times management$$

4) These two models above are fitted on the validation data, because deciding on the model with only training set would not be accurate. They are compared based on the same performance criteria. The adjusted R^2 values for both models dropped, compared to the values from the training set, but the amount of variation explained by the models are still adequate – 56.82% for lm3 and 57.77% for lm7. lm7 also performs slightly better on the test set on all criteria. The comparison can be seen in *Table 3*. Hence, the ultimate model of preference is lm7.

	Performance Criteria			
Model	**Res. Std.**	**Adj. R^2**	**AIC**	**PRESS**
lm3	0.1499	0.5682	-184.4799	4.6290
lm7	0.1486	0.5760	-188.1234	4.5595

Table 3: Performance measures of the selected models on the validation dataset

5) The preferred model, which is lm7, is fitted to the entire dataset. The model, with all coefficients being significant, is as follows:

$$\widehat{\sqrt{SWI}} = 0.6727 + 0.0720 \times SWF + 0.0780 \times SWF^2 + 0.0180 \times temp. +0.0229 \times manag.$$

	Performance Criteria			
Model	**Res. Std.**	**Adj. R^2**	**AIC**	**PRESS**
lm7train	0.1397	0.6889	-212.7907	3.9893
lm7test	0.1486	0.5760	-188.1234	4.5595
lm7full	0.1439	0.6385	-408.8606	8.3933

Table 4: Performance measures of the ultimate regression model on the datasets

According to the values above in *Table 4*, residual standard error and adjusted R^2 values for the model applied on the entire dataset are in between the values for the models applied on the training and test sets. Also, AIC and PRESS values are almost doubled, due to the sample size. All variables in the model have positive coefficients. SWF and SWF^2 have the highest coefficient. It is reasonable, because the correlation between SWI and SWF is the highest among the correlations. Also, as temperature increases, biodiversity increases, because the environmental conditions will be suitable to more species. Similarly, being subjected to nature management longer increases biodiversity.

Based on the model, the square-root of SWI is significantly influenced by SWF, the square of SWF, temperature and management. However, this time, interpreting the outcomes might be difficult, since transformation is applied. Back transformation is needed when the predictions are of interest. The extent to which the variables affect SWI are different. Based on the correlations between SWI and the variables (see *Table 1*), the most influential variable seems to be SWF (SWF^2) and the least influential variable seems to be temperature. However, based on the residual standard deviation reductions, the least influential variable is management. These values for the model combinations are given in *Table 5*. X_1 is SWF, X_2 is SWF^2, X_3 is temperature and X_4 is management.

Variables in the model	p	Res. Std.
X_1	1	0.1829
X_2	1	0.1807
X_3	1	0.2203
X_4	1	0.2276
X_1 , X_2	2	0.1808
X_1 , X_3	2	0.1568
X_1 , X_4	2	0.1734
X_2 , X_3	2	0.1548
X_2 , X_4	2	0.1713
X_3 , X_4	2	0.2078
X_2 , X_3 , X_4	3	0.1439
X_1 , X_3 , X_4	3	0.1458
X_1 , X_2 , X_4	3	0.1714
X_1 , X_2 , X_3	3	0.1548
X_1 , X_2 , X_3 , X_4	4	0.1397

Table 5: Residual std. of the models with different variables

- Temperature is more influential than management.
  - 0.2276 – 0.2078 > 0.2203 – 0.2078
  - 0.1734 – 0.1458 > 0.1568 – 0.1458
  - 0.1713 – 0.1439 > 0.1548 – 0.1439
  - 0.1714 – 0.1397 > 0.1548 – 0.1397

This approach is more accurate to compare the influence of temperature and management, because of the units of them. Even though one unit increase of management affects the response more than one unit increase of temperature (0.0229 > 0.0180), temperature changes in Celsius is easier to observe than management changes in years.

6) Now, the variables of interest are duration (response) and temperature (predictor). Since the relationship between these two variables is not strictly linear (see *Figure 2* below), a non-parametric model is needed. The analysis is done only on the training set.

(a) Six non-parametric models are fitted according to lowess method – with k = 1, 2 and s = 0.25, 0.50 and 0.75. The comparison table is given in *Table 6*. Based on the selected performance criteria, the best model seems to be nonp4, because it has the highest ENP (Equivalent Number of Parameters) value and the lowest error rates. Pairwise ANOVA comparisons reach the same result.

			Performance Criteria			
Model	**degree (k)**	**span (s)**	**ENP**	**Res. Std.**	**MSE**	**MAPE**
nonp1	1	0.25	8.34	2.0190	3.8424	0.0492
nonp2	1	0.50	4.75	2.1400	4.3309	0.0529
nonp3	1	0.75	3.39	2.3300	5.3070	0.0556
nonp4	2	0.25	13.86	2.0070	3.6879	0.0479
nonp5	2	0.50	7.70	2.0320	3.9390	0.0499
nonp6	2	0.75	5.37	2.0500	4.0695	0.0505

Table 6: Comparison of the non-parametric models

(b) Besides non-parametric models, quadratic linear models are fitted. Different combinations of quadratic terms are tried to select the best model. The comparison table is given in *Table 7*.

		Performance Criteria			
Model	**Orders of the temperature variable**	**Res. Std.**	**Adj. R^2**	**AIC**	**PRESS**
lmq1	2	3.471	0.4694	1069.3170	2574.9160
lmq2	1 , 2	2.536	0.7168	944.7579	1449.2120
lmq3	3	3.872	0.3397	1113.0480	3303.5480
lmq4	1 , 3	2.382	0.7501	919.7226	1262.8240
lmq5	2 , 3	2.157	0.7950	880.0773	973.7895
lmq6	1 , 2 , 3	2.102	0.8054	870.6939	902.9128

Table 7: Performance measures of the quadratic linear models

Based on the figures above, the best model seems to be lmq6, because it has the highest adjusted R^2 value and the lowest values for error criteria. The model is as follows:

$$\widehat{duration} = 24.1328 - 1.4956 \times temperature\ + 0.1917 \times temperature^2 - 0.0044 \times temperature^3$$

Main significant variable is the cubic term which has the lowest p-value. First and third order temperature coefficients seem to penalize the long duration, while second order temperature coefficient seems to reinforce it. However, the coefficients are not easy to interpret. They might even be considered as nonsense. Overfitting might be in consideration. Also, it does not satisfy Gauss-Markov conditions. Residuals are not correlated. The reason is that the model uses different orders of the same variable, hence the high correlation.

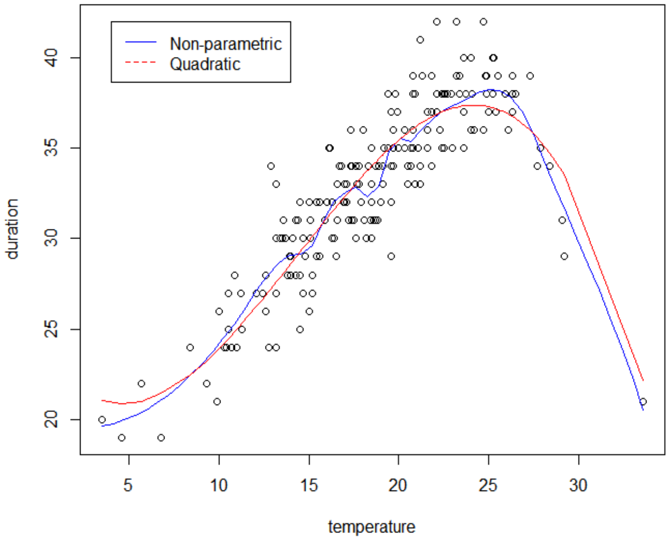

Figure 2: Fitted values for the both models

(c) The plot of the data with the fitted values for both models are given in *Figure 2*. According to visual interpretation, quadratic model seems to fit better, because it provides a smoother representation of the values. On the other hand, non-parametric fit line has some grits, as one might say it causes overfitting. Increasing the span of non-parametric model provides a smoother plot, but it may cause larger errors – there is a trade-off.

(d) In order to select the model that fits the data better, two models are compared based on certain prediction error criteria. The comparison could not be done with ANOVA, because non-parametric model does not assume any distribution and test statistics could not be obtained exactly. The comparison table is given below in *Table 8*.

	Performance Criteria			
Model	**Res. Std.**	**MSE**	**MAD**	**MAPE**
nonp4	2.0070	3.6879	1.5367	0.0479
lmq6	2.1020	4.3301	1.6768	0.0523

Table 8: Performance measures of the non-parametric and quadratic models

As can be seen above, non-parametric model has lower values on all performance criteria compared. Also, residuals of this model seem normally distributed – the p-value of the Shapiro-Wilk test is 0.8635. Hence, it can be concluded that the non-parametric model fits the data better. However, to reach more conclusive results, the models should be applied on the validation set as well.

Another suggestion might be to restrict the temperature parameter around 25°C, so that a strictly positive correlation might occur. This may also cause problems, because omitting certain observations might be misleading.

APPENDIX – R CODES

```
#Importing relevant libraries
library(car)
library(MASS)
library(robustbase)
library(MPV)

#Importing data
data.full = read.table(file.choose(),header=T)

#Dividing data into training and test sets
set.seed(0767950)
d.test <- sample(1:dim(data.full)[1], 200 )
data.test <- data.full[d.test, ]
data.training <- data.full[-d.test, ]

#1) EXPLORATORY DATA ANALYSIS
sum(is.na(data.training)) #0. No missing values.
summary(data.training)
attach(data.training)
boxplot(SWI, SWF, temperature, size, management, duration)
detach(data.training)
data.training.standardized <- data.frame(scale(data.training))
summary(data.training.standardized)
attach(data.training.standardized)
boxplot(SWI, SWF, temperature, size, management, duration)
detach(data.training.standardized)
#Standardized data is examined.
#No significant problem with outliers and skewness.

round(cor(data.training), 4)
#High positive correlations - SWI&SWF and temperature&duration

pairs(data.training) #temperature&duration pair shows some interesting properties.
plot(duration~temperature, data=data.training)
abline(v=25, col="blue")
#Irregularity on duration, duration variable is discarded for parametric analyses.

#2) FITTING A FIRST-ORDER LINEAR REGRESSION MODEL

#(a) Model fit
lm1 <- lm(SWI~SWF+temperature+size+management, data=data.training)
summary(lm1) #size isn't a significant variable.
par(mfrow = c(2,2))
plot(lm1) #There may be some problems.
par(mfrow = c(1,1))

#(b) Gauss-Markov conditions
gauss_markov <- function(linear_model) {
        #First Gauss-Markov condition. H0: E[residuals] = 0.
        gm1 <- t.test(linear_model$residuals, mu=0)
        #Second Gauss-Markov condition. H0: Homoscedasticity.
        gm2 <- ncvTest(linear_model)
        #Third Gauss-Markov condition. H0: Residuals are uncorrelated.
        gm3 <- durbinWatsonTest(linear_model)
        #All tests should have p-values above 0.05.
        results <- c(gm1$p.value > 0.05, gm2$p > 0.05, gm3$p > 0.05)
        if(sum(results) == 3) {
                return("All Gauss-Markov conditions are satisfied.")
                } else {
                print("Not all Gauss-Markov conditions are satisfied.")
                if(results[1] == F){
                        print("Residuals are not zero, on average.")
                        }
                if(results[2] == F){
                        print("Residuals are not homoscedastic.")
                        }
                if(results[3] == F){
                        print("Residuals are not uncorrelated.")
                        }
        }
}
gauss_markov(lm1) #Conditions aren't satisfied.

#(c) Multicollinearity
vifs <- vif(lm1)
sum(vifs > 10) #0
#No multicollinearity according to variation inflation factors.
```

```
rxx <- cor(data.training[,-c(1,6)])
eigen_rxx <- eigen(rxx)$values
condition <- sqrt(eigen_rxx[1]/eigen_rxx[2:4])
sum(condition > 30) #0
#No multicollinearity according to condition numbers.

#(d) Influential outliers
#Studentized residuals
lm1_studres <- studres(lm1)
plot(lm1_studres)
abline(h=c(-2.5,2.5))
#Only one influential point.

#DFFITS values
p=4; n=200
sum(abs(dffits(lm1)) > 2*sqrt(p/n))
#15 influential points.

#Cook's distances
sum(cooks.distance(lm1) > 1)
#No influential points.

#DFBETAS values
dfbetas <- (abs(dfbetas(lm1)) > 2/sqrt(n))[,2:5]
sum(rowSums(dfbetas) > 1)
#15 influential points.

#General influence table
influence.measures(lm1)
#3 influential points. (1,27,43)

#Robust diagnostics
lts1 <- ltsReg(SWI~SWF+temperature+size+management, data=data.training)
summary(lts1)
plot(lts1, which="rdiag")
#7 good leverage points and 1 vertical outlier.
#Vertical outlier is obs. 355.

#3) FINDING A BETTER MODEL
#Backward elimination
step_lm_backward <- stepAIC(lm1, direction = "back")
summary(step_lm_backward)
#Forward selection
step_lm_forward <- stepAIC(lm(SWI~1,data=data.training), direction = "forward", scope=formula(lm1))
summary(step_lm_forward)
#Stepwise regression
step_lm_stepwise <- stepAIC(lm(SWI~1,data=data.training), direction = "both", scope=formula(lm1))
summary(step_lm_stepwise)
#All 3 give the same model. (SWI ~ SWF + temperature + management)

#Updating the model
lm2 <- step_lm_stepwise
summary(lm2)
gauss_markov(lm2) #Conditions aren't satisfied.

#Creating new models using transformation
lm3 <- lm(sqrt(SWI)~SWF+temperature+management, data=data.training) #Square-root transformation
summary(lm3)
gauss_markov(lm3) #Conditions are satisfied.

lm4 <- lm(log(SWI)~SWF+temperature+management, data=data.training) #Logarithmic transformation
summary(lm4)
gauss_markov(lm4) #Conditions aren't satisfied.

lm_boxcox <- boxcox(lm1, lambda = seq(-2,2,0.001), plotit = TRUE) #Finding lambda for Box-Cox
transformation
best_lambda <- lm_boxcox$x[which(lm_boxcox$y == max(lm_boxcox$y))]
best_lambda #0.645
range(lm_boxcox$x[lm_boxcox$y > max(lm_boxcox$y)-qchisq(0.95,1)/2]) #Zero is outside the interval.
Lambda can be used.
data.training$boxcoxSWI <- (data.training$SWI^best_lambda-1)/best_lambda #Box-Cox transformation
lm5 <- lm(boxcoxSWI ~ SWF+temperature+management, data=data.training)
summary(lm5)
gauss_markov(lm5) #Conditions are satisfied.

lm6 <- lm(SWI ~ SWF + I(SWF^2)+temperature+management, data=data.training) #Square transformation
summary(lm6)
gauss_markov(lm6) #Conditions are satisfied.

lm7 <- lm(sqrt(SWI) ~ SWF + I(SWF^2)+temperature+management, data=data.training)
```

```r
summary(lm7)
gauss_markov(lm7) #Conditions are satisfied.

lm8 <- lm(boxcoxSWI ~ SWF + I(SWF^2)+temperature+management, data=data.training)
summary(lm8)
gauss_markov(lm8) #Conditions are satisfied.

#Weighted regression models

stdev <- lm(abs(residuals(lm3))~SWF+temperature+management, data=data.training)
w <- 1/stdev$fitted.values^2 #Weights
lm3w <- lm(sqrt(SWI)~SWF+temperature+management, data=data.training, weights=w)
summary(lm3w)
gauss_markov(lm3w) #Conditions are satisfied.

stdev <- lm(abs(residuals(lm5))~SWF+temperature+management, data=data.training)
w <- 1/stdev$fitted.values^2 #Weights
lm5w <- lm(boxcoxSWI ~ SWF+temperature+management, data=data.training, weights=w)
summary(lm5w)
gauss_markov(lm5w) #Conditions are satisfied.

stdev <- lm(abs(residuals(lm6))~SWF+I(SWF^2)+temperature+management, data=data.training)
w <- 1/stdev$fitted.values^2 #Weights
lm6w <- lm(SWI ~ SWF+I(SWF^2)+temperature+management, data=data.training, weights=w)
summary(lm6w)
gauss_markov(lm6w) #Conditions are satisfied.

stdev <- lm(abs(residuals(lm7))~SWF+I(SWF^2)+temperature+management, data=data.training)
w <- 1/stdev$fitted.values^2 #Weights
lm7w <- lm(sqrt(SWI) ~ SWF+I(SWF^2)+temperature+management, data=data.training, weights=w)
summary(lm7w)
gauss_markov(lm7w) #Conditions are satisfied.

stdev <- lm(abs(residuals(lm8))~SWF+I(SWF^2)+temperature+management, data=data.training)
w <- 1/stdev$fitted.values^2 #Weights
lm8w <- lm(boxcoxSWI ~ SWF+I(SWF^2)+temperature+management, data=data.training, weights=w)
summary(lm8w)
gauss_markov(lm8w) #Conditions are satisfied.

#TEN MODELS TO COMPARE
AIC(lm3); AIC(lm5); AIC(lm6); AIC(lm7); AIC(lm8) #Unweighted AIC
AIC(lm3w); AIC(lm5w); AIC(lm6w); AIC(lm7w); AIC(lm8w) #Weighted AIC
PRESS(lm3);PRESS(lm5); PRESS(lm6); PRESS(lm7); PRESS(lm8) #Unweighted PRESS
PRESS(lm3w);PRESS(lm5w); PRESS(lm6w); PRESS(lm7w); PRESS(lm8w) #Weighted PRESS

#CONCLUSION: BEST MODELS ARE lm3 & lm7.
#lm3: sqrt(SWI)~I(SWF^2)+temperature+management
#lm7: sqrt(SWI)~SWF+temperature+management
summary(lm3)
summary(lm7)

#4) FITTING MODELS TO VALIDATION DATA
lm3_val <- lm(sqrt(SWI)~SWF+temperature+size+management, data=data.test)
summary(lm3_val)
gauss_markov(lm3_val)

lm7_val <- lm(sqrt(SWI)~SWF+I(SWF^2)+temperature+management, data=data.test)
summary(lm7_val)
gauss_markov(lm7_val)

#Performance measures
AIC(lm3_val); AIC(lm7_val)
PRESS(lm3_val); PRESS(lm7_val)

#CONCLUSION: lm7 IS (SLIGHTLY) BETTER.

#5) FITTING THE ULTIMATE MODEL ON THE FULL DATA
lm7_full <- lm(sqrt(SWI)~SWF+I(SWF^2)+temperature+management, data=data.full)
summary(lm7_full)
gauss_markov(lm7_full)
AIC(lm7_full); PRESS(lm7_full)

# Model combinations for partial contributions
lm7_w_swf <- lm(sqrt(SWI)~SWF, data=data.full); summary(lm7_w_swf)
lm7_w_swf2 <- lm(sqrt(SWI)~I(SWF^2), data=data.full); summary(lm7_w_swf2)
lm7_w_temp <- lm(sqrt(SWI)~temperature, data=data.full); summary(lm7_w_temp)
lm7_w_mgmt <- lm(sqrt(SWI)~management, data=data.full); summary(lm7_w_mgmt)
lm7_ww1 <- (lm(sqrt(SWI)~SWF+I(SWF^2), data=data.full)); summary(lm7_ww1)
lm7_ww2 <- (lm(sqrt(SWI)~SWF+temperature, data=data.full)); summary(lm7_ww2)
lm7_ww3 <- (lm(sqrt(SWI)~SWF+management, data=data.full)); summary(lm7_ww3)
lm7_ww4 <- (lm(sqrt(SWI)~I(SWF^2)+temperature, data=data.full)); summary(lm7_ww4)
```

```
lm7_ww5 <- (lm(sqrt(SWI)~I(SWF^2)+management, data=data.full)); summary(lm7_ww5)
lm7_ww6 <- (lm(sqrt(SWI)~temperature+management, data=data.full)); summary(lm7_ww6)
lm7_wo_swf <- lm(sqrt(SWI)~I(SWF^2)+temperature+management, data=data.full); summary(lm7_wo_swf)
lm7_wo_swf2 <- lm(sqrt(SWI)~SWF+temperature+management, data=data.full); summary(lm7_wo_swf2)
lm7_wo_temp <- lm(sqrt(SWI)~SWF+I(SWF^2)+management, data=data.full); summary(lm7_wo_temp)
lm7_wo_mgmt <- lm(sqrt(SWI)~SWF+I(SWF^2)+temperature, data=data.full); summary(lm7_wo_mgmt)

#CONCLUSION: Variable importance is SWF^2 > SWF > temperature > management

#6) NONPARAMETRIC AND QUADRATIC MODELS

#(a) Non-parametric model
attach(data.training)
nonp1 <- loess(duration~temperature, span=0.25, degree=1); nonp1
nonp2 <- loess(duration~temperature, span=0.5, degree=1); nonp2
nonp3 <- loess(duration~temperature, span=0.75, degree=1); nonp3
nonp4 <- loess(duration~temperature, span=0.25, degree=2); nonp4
nonp5 <- loess(duration~temperature, span=0.5, degree=2); nonp5
nonp6 <- loess(duration~temperature, span=0.75, degree=2); nonp6

#Comparing with the null model
anova(nonp1, update(nonp1,span=1))
anova(nonp2, update(nonp2,span=1))
anova(nonp3, update(nonp3,span=1))
anova(nonp4, update(nonp4,span=1))
anova(nonp5, update(nonp5,span=1))
anova(nonp6, update(nonp6,span=1))
#All models are better than the null model.

#MSE and MAPE
mean(nonp1$residuals^2); mean(abs(nonp1$residuals)/duration)
mean(nonp2$residuals^2); mean(abs(nonp2$residuals)/duration)
mean(nonp3$residuals^2); mean(abs(nonp3$residuals)/duration)
mean(nonp4$residuals^2); mean(abs(nonp4$residuals)/duration)
mean(nonp5$residuals^2); mean(abs(nonp5$residuals)/duration)
mean(nonp6$residuals^2); mean(abs(nonp6$residuals)/duration)

#CONCLUSION: BEST MODEL IS nonp4.

#(b) Quadratic linear model
lmq1 <- lm(duration~I(temperature^2), data=data.training); summary(lmq1); gauss_markov(lmq1)
lmq2    <-    lm(duration~temperature+I(temperature^2),    data=data.training);    summary(lmq2);
gauss_markov(lmq2)
lmq3 <- lm(duration~I(temperature^3), data=data.training); summary(lmq3); gauss_markov(lmq3)
lmq4    <-    lm(duration~temperature+I(temperature^3),    data=data.training);    summary(lmq4);
gauss_markov(lmq4)
lmq5   <-   lm(duration~I(temperature^2)+I(temperature^3),   data=data.training);   summary(lmq5);
gauss_markov(lmq5)
lmq6    <-    lm(duration~temperature+I(temperature^2)+I(temperature^3),    data=data.training);
summary(lmq6); gauss_markov(lmq6)
#Performance criteria of the quadratic models
AIC(lmq1);AIC(lmq2);AIC(lmq3);AIC(lmq4);AIC(lmq5);AIC(lmq6)
PRESS(lmq1);PRESS(lmq2);PRESS(lmq3);PRESS(lmq4);PRESS(lmq5);PRESS(lmq6)

#CONCLUSION: THE BEST MODEL IS lmq6.

#(c) Plot and fits
plot(duration~temperature)
abline(v=25, lty=3)
lines(loess.smooth(temperature, duration, span=0.25, degree=2), col="blue") #Non-parametric fit
lines(temperature,   24.132803-1.495634*temperature+0.191738*temperature^2-0.004434*temperature^3,
col="red") #Quadratic fit
legend(4,42, legend=c("Non-parametric", "Quadratic"), col=c("blue","red"), lty=1:2)

#Checking the validity of the non-parametric model
nonp_null <- update(nonp4, span=1)
scatter.smooth(temperature, residuals(nonp4), span=0.25, degree=2) #No problem. Around 0.
scatter.smooth(temperature, residuals(nonp4), span=1, degree=1) #No problem. Around 0.
qqnorm(residuals(nonp4)); qqline(residuals(nonp4)) #Seems normal.
shapiro.test(residuals(nonp4)) #p-value 0.8635
abline(h=0, lty=2)

#(d) Non-parametric vs. Quadratic comparison test
#Res.Std., MSE, MAD, MAPE
nonp4; summary(lmq6)
mean(nonp4$residuals^2); mean(lmq6$residuals^2)
mean(abs(nonp4$residuals)); mean(abs(lmq6$residuals))
mean(abs(nonp4$residuals)/duration); mean(abs(lmq6$residuals)/duration)
```