Machine Learning Exam March 2020

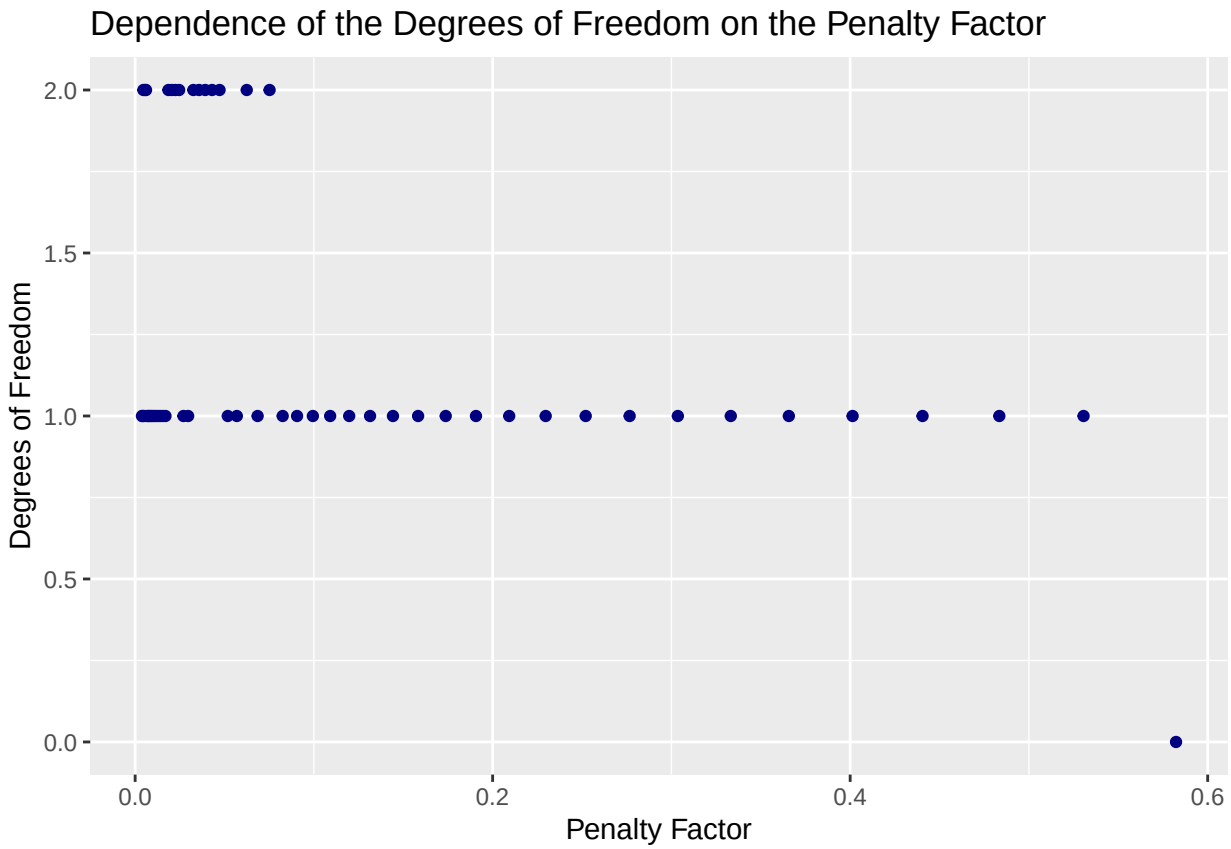
Christophoros Spyretos

Assignment 1

```
# import data
temperature <- read.csv2("Dailytemperature.csv")

# creating new features
x <- temperature$Day

phi1 <- NULL
phi2 <- NULL
k <- seq(-50,50,1)

for (i in k){
  phi1 <- sin(0.5^(i)*x)
  phi2 <- cos(0.5^(i)*x)
}

temperature$Phi1 <- phi1
temperature$Phi2 <- phi2
```

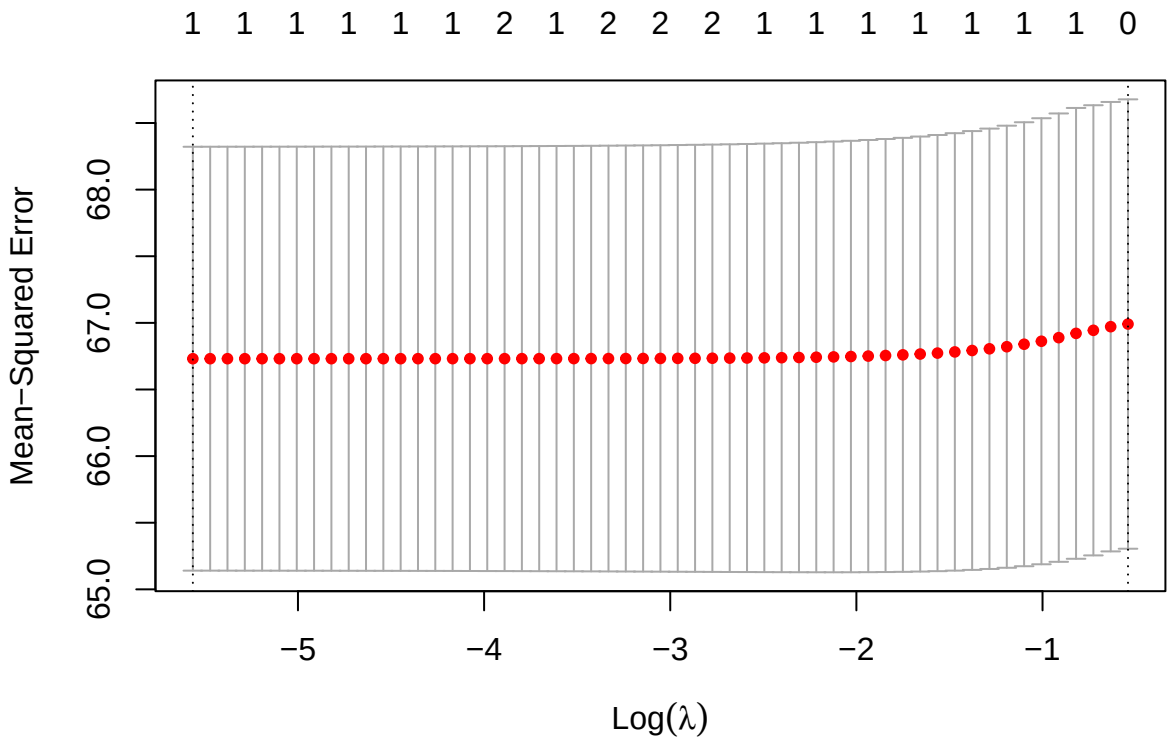
```
# fitting lasso regression model
X <- temperature[,-2]
Y <- temperature[,2]

lasso <- glmnet(as.matrix(X), as.matrix(Y), alpha=1)

# summary(lasso)
```

```
# plot of the dependence of the degrees of freedom on the value of the penalty factor
df_plot <- data.frame( "df" <- lasso$df,
                       "lambda" <- lasso$lambda)

ggplot( data = df_plot, aes(x=lambda,y=df)) +
  geom_point( color = "navy") +
  labs(x="Penalty Factor",
       y= "Degrees of Freedom",
       title = "Dependence of the Degrees of Freedom on the Penalty Factor")
```



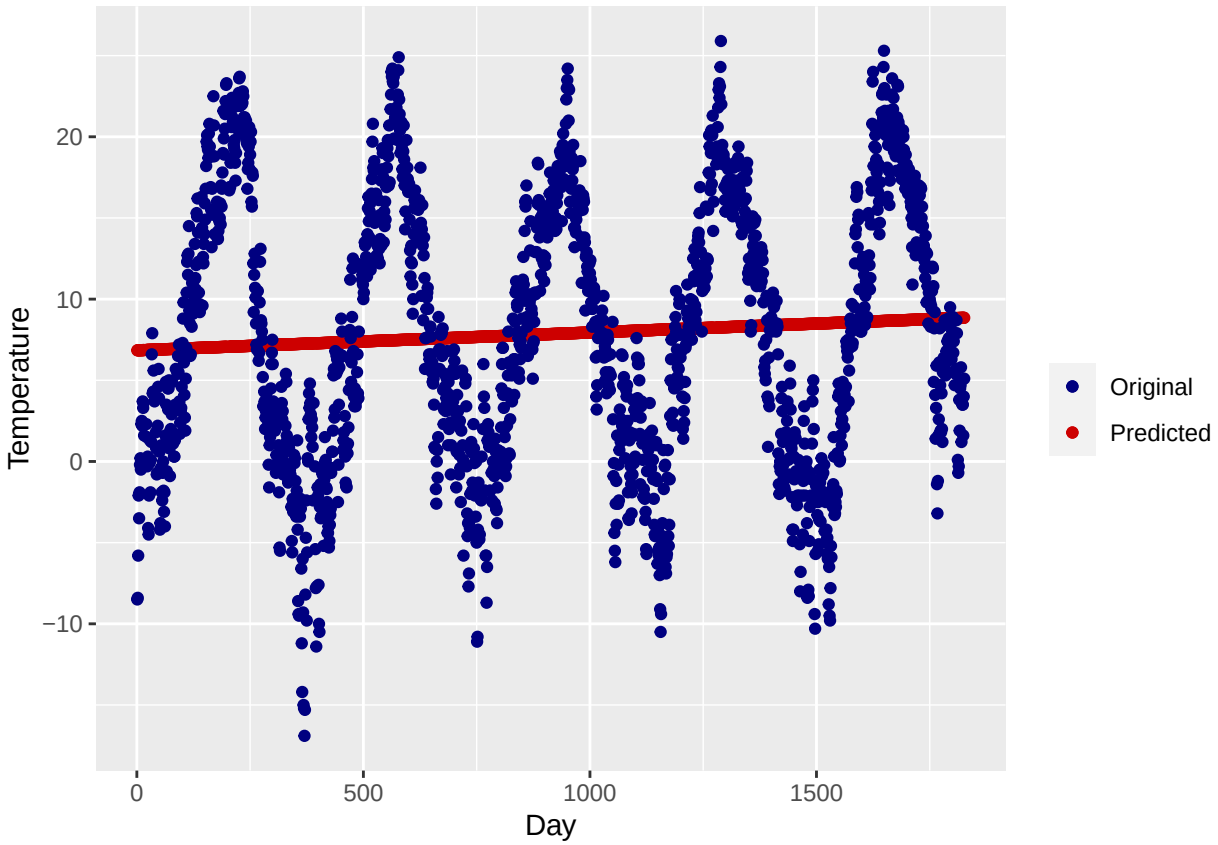
From the above plot it could be assumed that there is a downward tendency between the *Degrees of Freedom* and the *Penalty Factor*.

```
# cross-validation
cv_lasso <- cv.glmnet(as.matrix(X), as.matrix(Y), alpha = 1)

plot(cv_lasso)
```



The optimal lambda , which equals 0.003831144, and the $\log(\lambda) = -4$ have similar MSE values. The optimal lambda has a shred higher MSE, but due to the similar confidence bounce, there is no evident assumption if it is statistically significant.

```
# number of non-zero features corresponding to the optimal penalty factor
amount <- sum(coef(lasso, s = cv_lasso$lambda.min)[,1] != 0)

# type = "link" returns the fitted values
# type = "response" gives the same output "link" for "gaussian" family
# type = "coefficients" returns the model codfficients
# type = "nonzero' retuns a list of the indices of the nonzero coefficients for each value of s.

# # calculating fitted data corresponding to the optimal penalty factor
fitted_values <- predict(cv_lasso , as.matrix(temperature[,-2]) , type = "link",
                         s=cv_lasso$lambda.min)

# data for the plot
optimal_df <- data.frame( "Original"= temperature$Temperature,
                          "Predicted" = fitted_values,
                          "Day" = temperature$Day)

# time series plot of the original and the fitted data corresponding to the optimal penalty factor
my_scatterplot <-ggplot(optimal_df,aes(x=s1, y= Original)) +
  geom_point(aes(x=Day, y= s1, color = "red3")) +
  geom_point(aes(x=Day, y= Original, color = "navy")) +
  labs(x = "Day",y = "Temperature") +
  scale_color_manual(values=c("navy","red3"),
                     name = "",
                     labels = c("Original","Predicted"))

my_scatterplot
```



From the above plot, it could be assumed that the model's predictions are badly. It could be seen that the predicted values are only positive numbers, whereas the actual values are both positive and negative. However, it could be noticed that as the days pass, the winter seasons have become less cold and the summer seasons hotter.

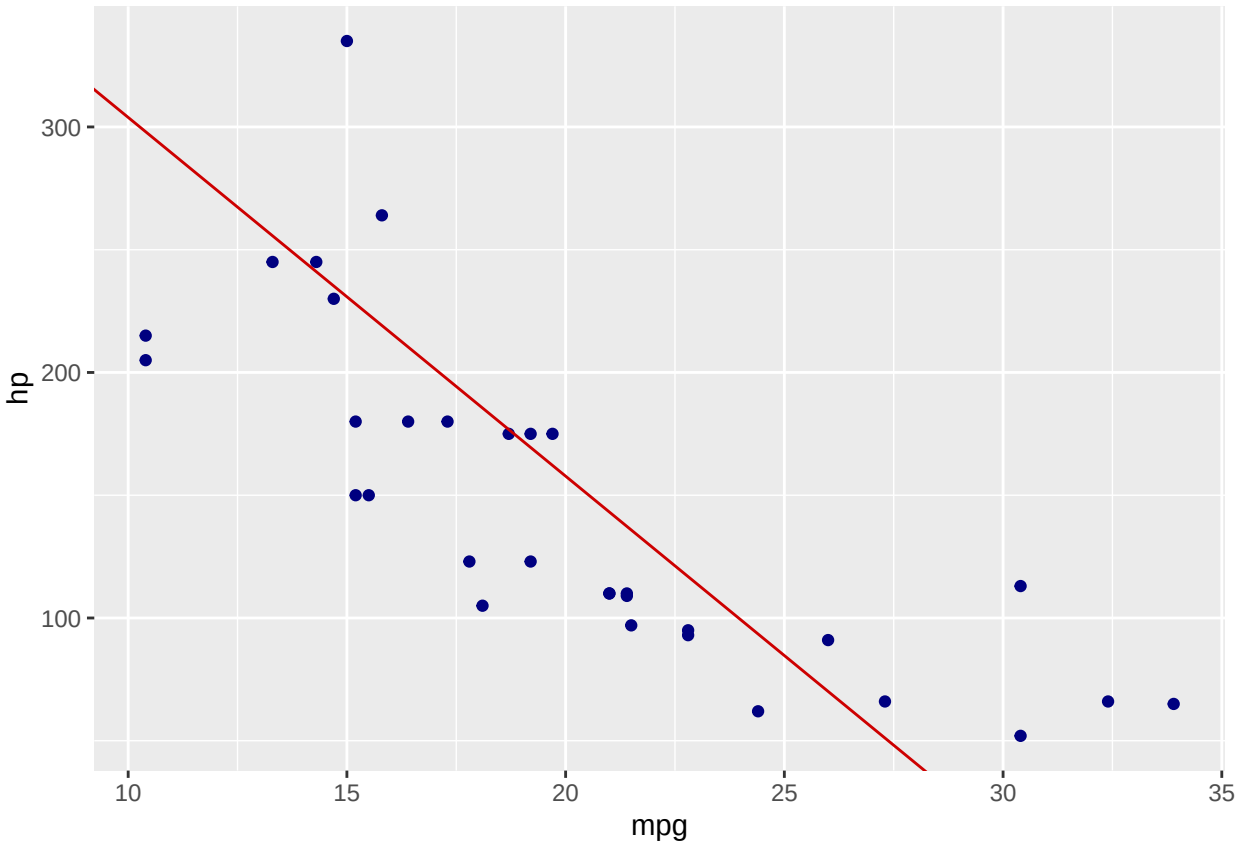
Assignment 2

```
mtcars <- mtcars

#preparing data
mtcars <- mtcars[,-c(2,3,5:11)]
```

```
cov_matrix <- cov(mtcars)
eig <- eigen(cov_matrix)
eigen_vec <- eig$vectors
eigen_val <- eig$values

cat("The components of the first principle component are", eigen_vec[,1],".")
```

```
## The components of the first principle component are -0.06827783 0.9976663 .
```

```
ggplot(data = mtcars, aes(x = mpg, y = hp)) +
  geom_point(color = "navy") +
  geom_abline( slope = eigen_vec[2,1]/eigen_vec[1,1],
               intercept = 450, color = "red3")
```



From the above plot, it is evident that the line demonstrates a negative relationship between the two values. Moreover, it is expected because, generally, more horsepower means fewer miles per gallon.