# MBA 6693 Business Analytics

Assignment 3: Cross Validation

*Name: Vinayak B. Menon*

*01/08/2020*

**Objective:**

This report aims to create a model that establishes a relationship between the net hourly electrical energy output of a Combined Cycle Power Plant based on the hourly average values of 4 variables, namely, Ambient Temperature (AT), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V). We will first perform exploratory data analysis, after which we provide possible models based on regression. We end the report with an analysis of the model performance using various metrics.

A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load, and is included as an excel document.

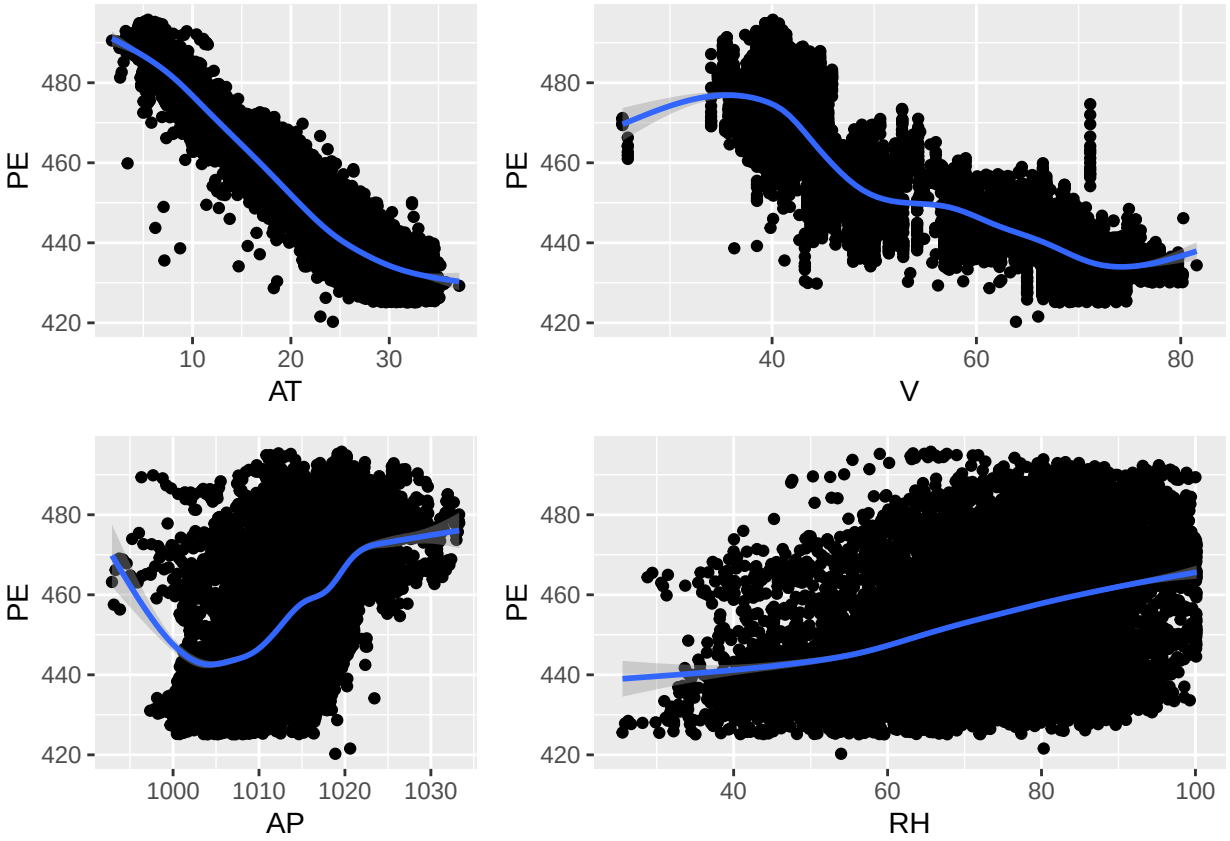
The code for the data analysis has been provided in the appendix.

**Exploratory Data Analysis:**

We first present the summary statistics of the dataset:

```
##        AT              V               AP              RH        
##  Min.   : 1.81   Min.   :25.36   Min.   : 992.9   Min.   : 25.56  
##  1st Qu.:13.51   1st Qu.:41.74   1st Qu.:1009.1   1st Qu.: 63.33  
##  Median :20.34   Median :52.08   Median :1012.9   Median : 74.97  
##  Mean   :19.65   Mean   :54.31   Mean   :1013.3   Mean   : 73.31  
##  3rd Qu.:25.72   3rd Qu.:66.54   3rd Qu.:1017.3   3rd Qu.: 84.83  
##  Max.   :37.11   Max.   :81.56   Max.   :1033.3   Max.   :100.16  
##        PE        
##  Min.   :420.3  
##  1st Qu.:439.8  
##  Median :451.6  
##  Mean   :454.4  
##  3rd Qu.:468.4  
##  Max.   :495.8
```

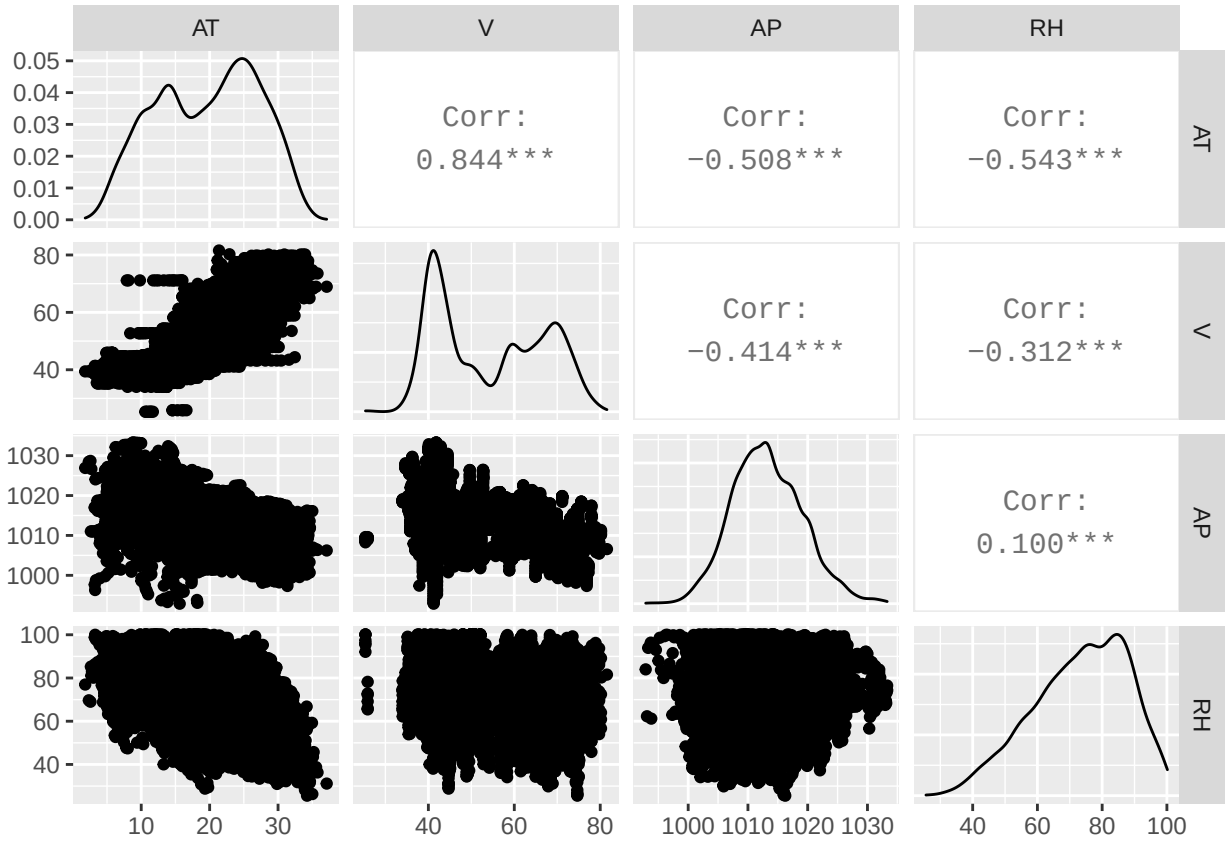
As we can see, all the features are continuous in nature, and there is no missing data present in the data as well. However, normalizing will be required to scale the variables to a uniform format. Before this, it will be useful to present some plots to describe some relationships between the dependent and independent variables and between the independent variables.




Based on the scatterplots, we can see that there is a visible and strong negative relationship between the net hourly electrical output of the power plant and the Ambient Temperature feature, implying that the greater the ambient temperature, the less output the turbines produce. There is also a relatively visible negative relationship between the hourly net electrical output of the power plant with the exhaust vacuum feature as well.

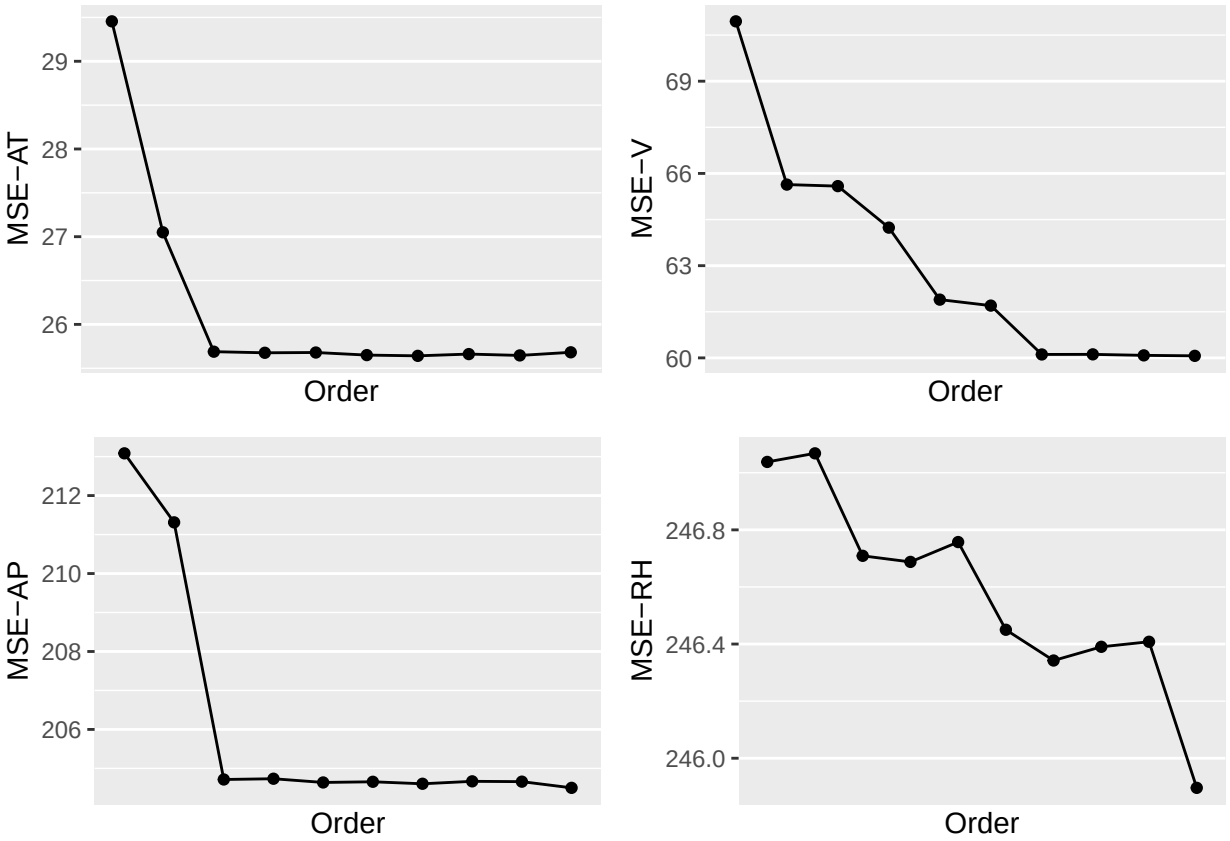
We can also see a slight positive relationship to the electrical output in the scatter plots for relative humidity and ambient pressure. Note that the relationship is also riddled with a higher degree of outliers.

Now we move on to a bivariate analysis of the independent variables, to account for any interaction effects.




Based on the above correlogram, we can see that there is a strong correlation observed for AT with all othe other features, while relatively mild relationship between exhaust vacuum and both ambient pressure and relative humidity. This indicates a potential need to include interaction effects.

Before we move on to normalizing the dataset, it may be helpful to know whether there are any non linear relationships between the dependent and independent variables. We can compare the MSE values of a regression model with an increase in order of the single dependent variable.




Above we have plotted the MSE values of regression models with the dependent variable PE and the corresponding independent variable under multiple orders. Based on the plots, we can see that there is generally a very small level of decrease in the MSE values as we move the order of the terms from 1 to 10. But there is a notable steep decrease in MSE for AT and AP while V and RH tend to have a less steep and quite jagged decrease. Although the decrease is quite low, including polynomial terms may help improve the model performance, especially for AP and AT features. Note that we will restrict our use of polynomial order terms to a maximum of order 3.

Now we can proceed to creating a few models, after providing a summary of the normalized dataset.

```
##        AT                 V                 AP                 RH
##  Min.   :-2.39400   Min.   :-2.2778   Min.   :-3.42984   Min.   :-3.2704
##  1st Qu.:-0.82405   1st Qu.:-0.9888   1st Qu.:-0.70032   1st Qu.:-0.6836
##  Median : 0.09309   Median :-0.1752   Median :-0.05373   Median : 0.1141
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.81433   3rd Qu.: 0.9627   3rd Qu.: 0.67369   3rd Qu.: 0.7891
##  Max.   : 2.34268   Max.   : 2.1447   Max.   : 3.37458   Max.   : 1.8391
##        PE
##  Min.   :-1.9983
##  1st Qu.:-0.8563
##  Median :-0.1649
##  Mean   : 0.0000
##  3rd Qu.: 0.8241
##  Max.   : 2.4254
```

**Modelling:**

We form 4 models to establish the relationship between the net hourly electrical output of the power plant and the various features explained apriori. The first model includes all of the variables as is, while the second model includes higher order terms as well (again, we restrict the order to 3 to avoid over complexity).

The third model includes interaction effects between the independent variables and the final model includes both interaction effects and higher order effects. We present a summary of the models below, with the required changes in terms where needed.

Model 1:

```
##
## Call:
## lm(formula = PE ~ ., data = ccpp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5450 -0.1855 -0.0069  0.1875  1.0416
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.280e-15  2.730e-03    0.000        1
## AT          -8.635e-01  6.676e-03 -129.342  < 2e-16 ***
## V           -1.742e-01  5.422e-03  -32.122  < 2e-16 ***
## AP           2.160e-02  3.291e-03    6.564 5.51e-11 ***
## RH          -1.352e-01  3.566e-03  -37.918  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2671 on 9563 degrees of freedom
## Multiple R-squared:  0.9287, Adjusted R-squared:  0.9287
## F-statistic: 3.114e+04 on 4 and 9563 DF,  p-value: < 2.2e-16
```

Model 2:

```
##
## Call:
## lm(formula = PE ~ . + I(AT^3) + I(AP^3), data = ccpp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54601 -0.18313 -0.00717  0.18543  1.03747
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.003177   0.002760   1.151     0.25
## AT          -0.895174   0.009702 -92.268  < 2e-16 ***
## V           -0.165988   0.005674 -29.254  < 2e-16 ***
## AP           0.036989   0.004869   7.598 3.30e-14 ***
## RH          -0.134927   0.003571 -37.784  < 2e-16 ***
## I(AT^3)      0.013229   0.002680   4.936 8.11e-07 ***
```

```
## I(AP^3)     -0.005176   0.001067  -4.850 1.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2664 on 9561 degrees of freedom
## Multiple R-squared:  0.9291, Adjusted R-squared:  0.929
## F-statistic: 2.087e+04 on 6 and 9561 DF,  p-value: < 2.2e-16
```

Model 3:

```
##
## Call:
## lm(formula = PE ~ . + (AT * AP) + (AT * V) + (AT * RH), data = ccpp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65900 -0.16798  0.00065  0.17414  1.08631
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -0.091034   0.003989  -22.823   <2e-16 ***
## AT          -0.791380   0.006964 -113.633   <2e-16 ***
## V           -0.226341   0.005610  -40.347   <2e-16 ***
## AP           0.042865   0.003358   12.764   <2e-16 ***
## RH          -0.104833   0.003498  -29.971   <2e-16 ***
## AT:AP        0.030273   0.003030    9.992   <2e-16 ***
## AT:V         0.108229   0.004441   24.370   <2e-16 ***
## AT:RH       -0.027743   0.003030   -9.156   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2529 on 9560 degrees of freedom
## Multiple R-squared:  0.9361, Adjusted R-squared:  0.9361
## F-statistic: 2.001e+04 on 7 and 9560 DF,  p-value: < 2.2e-16
```

Model 4:

```
##
## Call:
## lm(formula = PE ~ ., data = ccpp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5450 -0.1855 -0.0069  0.1875  1.0416
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.280e-15  2.730e-03    0.000        1
## AT          -8.635e-01  6.676e-03 -129.342  < 2e-16 ***
## V           -1.742e-01  5.422e-03  -32.122  < 2e-16 ***
## AP           2.160e-02  3.291e-03    6.564 5.51e-11 ***
```

```
## RH          -1.352e-01  3.566e-03  -37.918  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2671 on 9563 degrees of freedom
## Multiple R-squared:  0.9287, Adjusted R-squared:  0.9287
## F-statistic: 3.114e+04 on 4 and 9563 DF,  p-value: < 2.2e-16
```

Note that the order terms are restricted to those independent variables that showed a simple and visible decrease in MSE within upto 3 order terms. Since the decrease in MSE for all independent variables is feeble for upto 10 orders, it would be pointless to include any further.
For interaction effects, we have restricted ourselves to the interactions between independent variables whose correlations are above 0.5.
The final model includes the terms from all the models before. Since the p-values lie well below 0.05 for all the terms in each of the models, no changes are needed to the models. We will observe the model performance in the next section, where we also make our final inferences.

**Model Testing and Inference:**

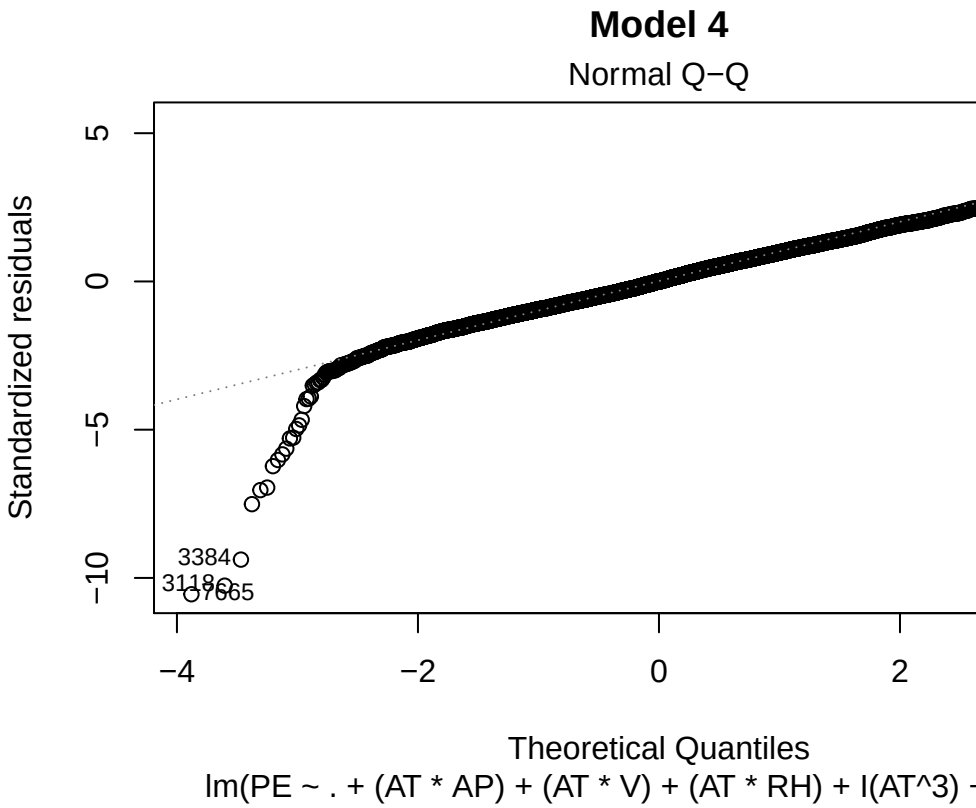
We publish the measures of fitness for the models listed in a tabular form below:

```
##           RSQ   RSE      AIC      BIC   MSE
## Model 1 0.929 0.267 1896.594 1939.591 0.071
## Model 2 0.929 0.266 1849.568 1906.897 0.071
## Model 3 0.936 0.253  852.216  916.711 0.064
## Model 4 0.936 0.252  816.818  895.646 0.064
```

A lot can be inferred from the above table. The main inference we can gather is that Model 4, i.e the model with both higher order terms and interaction effects, seems the best model out of the 4.

Based on the Adjusted R-squared values, Model 3 and Model 4 explain a higher amount of variance as compared to models 1 and 2. There is also a significant decrease in the AIC and BIC values of Models 1 and 2 when compared to Models 3 and 4. Although applying BIC does include some penalty for the extra terms included in Models 2,3 and especially Model 4, the end results do not differ. The values of MSE also produce similar results, where Models 3 and 4 are lower than that of Models 1 and 2. Thus we can safely reject the first and second models based on all the above criteria.

When comparing for Models 3 and 4, we can see that the RSE, AIC and BIC indicates that Model 4 is much more suitable in the end. In the end, we see that the interaction effects improve the model significantly, while the order terms provide a lesser increase in model fit (as can be seen when comparing Model 3 and Model 4).




We finally plot the QQ plot for Model 4:
The plot seems to indicate that the residuals tend to form a more heavy tailed distribution as compared to a normal distribution. A suitable distribution fit with higher kurtosis should solve this issue.

**Conclusion:**

We can conclude from the above analysis that interaction effects between ambient temperature and relative humidity, ambient pressure and exhaust vacuum significantly improves the model performance when compared to a standard multiple linear regression model. The addition of higher order terms improves the performance, though not as significantly as interaction effects. Combining interaction effects and higher order terms of ambient temperature and ambient pressure provides the best model out of the 4 to predict the net hourly electrical output of the power plant.

**Code:**

```
rm(list=ls())
#setting working directory
setwd("C:/Users/Vinayak/Documents/GitHub/Assignment A03")
#loading required packages
library(ggplot2)
library(GGally)
library(dplyr)
library(readxl)
library(broom)
library(boot)
library(gridExtra)
```

```r
#accessing CCPP data
ccpp <- read_xlsx("Folds5x2_pp.xlsx")
ccpp <- as.data.frame(ccpp)
attach(ccpp)


######################
# MODEL EXPLORATION #
######################

#summary of data
summary(ccpp)

#scatter plots
ATPE <- ggplot(ccpp,aes(AT,PE))+
         geom_point()+
         geom_smooth()

VPE<- ggplot(ccpp,aes(V,PE))+
        geom_point()+
        geom_smooth()

APPE<-ggplot(ccpp,aes(AP,PE))+
        geom_point()+
        geom_smooth()

RHPE<- ggplot(ccpp,aes(RH,PE))+
        geom_point()+
        geom_smooth()

grid.arrange(ATPE,VPE,APPE,RHPE, ncol = 2,nrow=2, widths = c(4, 6))
```

```r
#correlogram plot
ggpairs(ccpp[,-5])
```

```r
#plotting MSE values for various orders
cv_err <- matrix(0,nrow=10,ncol=4)

for (i in 1:10){
  glm_fit <- glm(PE~poly(AT,i), data = ccpp)
  cv_err[i,1] <- cv.glm(ccpp, glm_fit,K=10)$delta[1]

  glm_fit <- glm(PE~poly(V,i), data = ccpp)
  cv_err[i,2] <- cv.glm(ccpp, glm_fit,K=10)$delta[1]

  glm_fit <- glm(PE~poly(AP,i), data = ccpp)
  cv_err[i,3] <- cv.glm(ccpp, glm_fit,K=10)$delta[1]

  glm_fit <- glm(PE~poly(RH,i), data = ccpp)
  cv_err[i,4] <- cv.glm(ccpp, glm_fit,K=10)$delta[1]
}
cv_err <-as.data.frame(cv_err)
colnames(cv_err)<-c("AT","V","AP","RH")
```

```
cv_err$order<-1:10

#plotting MSE of each independent variable for orders upto 10
AT_cv<-ggplot(cv_err,aes(x=order,y=AT))+
  geom_path()+
  geom_point()+scale_x_discrete(name="Order",breaks=1:10)+scale_y_continuous(name = "MSE-AT")

V_cv<-ggplot(cv_err,aes(x=1:10,y=V))+
  geom_path()+
  geom_point()+scale_x_discrete(name="Order",breaks=1:10)+scale_y_continuous(name = "MSE-V")

AP_cv <-ggplot(cv_err,aes(x=1:10,y=AP))+
  geom_path()+
  geom_point()+scale_x_discrete(name="Order",breaks=1:10)+scale_y_continuous(name = "MSE-AP")

RH_cv<- ggplot(cv_err,aes(x=1:10,y=RH))+
  geom_path()+
  geom_point()+scale_x_discrete(name="Order",breaks=1:10)+scale_y_continuous(name = "MSE-RH")

grid.arrange(AT_cv,V_cv,AP_cv,RH_cv,ncol=2,nrow=2)
```

```
#normalizing data
for (i in 1:5) {
  ccpp[,i] <- (ccpp[,i]-mean(ccpp[,i]))/sd(ccpp[,i])
}
#summary of normalized data
summary(ccpp)

###################
# MODEL CREATION #
###################

#MODEL 1- STANDARD MULTIPLE LINEAR REGRESSION
m_1 <- lm(PE~.,data=ccpp)
summary(m_1)

#MODEL 2-HIGHER ORDER TERMS
m_2 <- lm(PE~.+I(AT^3)+I(AP^3),data=ccpp)
summary(m_2)

#MODEL 3- INTERACTION EFFECTS
m_3 <- lm(PE~.+(AT*AP)+(AT*V)+(AT*RH),data=ccpp)
summary(m_3)

#MODEL 4- INTERACTION EFFECTS +HIGHER ORDER TERMS
m_4 <- lm(PE~.+(AT*AP)+(AT*V)+(AT*RH)+I(AT^3)+I(AP^3),data=ccpp)
summary(m_1)

###################
# MODEL ANALYSIS #
###################

#Obtaining various measures of fit
```

```r
fit<-as.data.frame(matrix(0,ncol = 5,nrow = 4))
rownames(fit)<-c("Model 1","Model 2","Model 3","Model 4")
colnames(fit)<-c("RSQ","RSE","AIC","BIC","MSE")

#R-squared
fit$RSQ[1] <-summary(m_1)$adj.r.squared
fit$RSQ[2] <-summary(m_2)$adj.r.squared
fit$RSQ[3] <-summary(m_3)$adj.r.squared
fit$RSQ[4] <-summary(m_4)$adj.r.squared

#RSE
fit$RSE[1] <-summary(m_1)$sigma
fit$RSE[2] <-summary(m_2)$sigma
fit$RSE[3] <-summary(m_3)$sigma
fit$RSE[4] <-summary(m_4)$sigma

#AIC
fit$AIC[1] <-AIC(m_1)
fit$AIC[2] <-AIC(m_2)
fit$AIC[3] <-AIC(m_3)
fit$AIC[4] <-AIC(m_4)

#BIC
fit$BIC[1] <-BIC(m_1)
fit$BIC[2] <-BIC(m_2)
fit$BIC[3] <-BIC(m_3)
fit$BIC[4] <-BIC(m_4)

#MSE through Cross Validation
fit$MSE[1]<-cv.glm(ccpp,glm(PE~.,data=ccpp),K=10)$delta[1]
fit$MSE[2]<-cv.glm(ccpp,glm(PE~.+I(AT^3)+I(AP^3),data=ccpp),K=10)$delta[1]
fit$MSE[3]<-cv.glm(ccpp,glm(PE~.+(AT*AP)+(AT*V)+(AT*RH),data=ccpp),K=10)$delta[1]
fit$MSE[4]<-cv.glm(ccpp,glm(PE~.+(AT*AP)+(AT*V)+(AT*RH)+I(AT^3)+I(AP^3),data=ccpp),K=10)$delta[1]

#printing measure of fit table
print(round(fit,3))

#QQ PLOT OF MODEL 4
plot(m_4,2,main ="Model 4")
```