Moiz Kapadia **CEE322** **Lab 3**

PART 1

Introduction

The efficient operation of a combined cycle plant can provide environmental benefits in terms of pollution reduction, as well as economic benefits to owners and customers. Through confirmatory data analysis, average hourly plant conditions can show which parameters relate to power output. The data set used in this analysis has 9,568 observations, which is roughly more than one year of operation.

	AT	V	AP	RH	PE
1	14.96	41.76	1024.07	73.17	463.26
2	25.18	62.96	1020.04	59.08	444.37

Figure 1. Snapshot of the dataset, which contains the operating conditions recorded as hourly averages over a handful of years. T = Temperature (degC), AP = Ambient Pressure (millibar), RH = Relative Humidity (RH), V = Exhaust Vacuum (cm Hg), PE = Net energy output (MW)

Temperature effects on Power Output

The first model tested postulates that there is a relationship between Power Output and Temperature. In terms of R, this relationship is expressed as the following:

$$\textit{Power Output} \sim \textit{Temperature} \quad [1]$$

A linear regression test confirmed that an increase in Temperature decreases the Power Output of the plant, with an R^2 value of 0.8989. This means that 89.89% of the variation in Power Output can be explained by Temperature. Figure 2 shows the regression line imposed on top of the data.

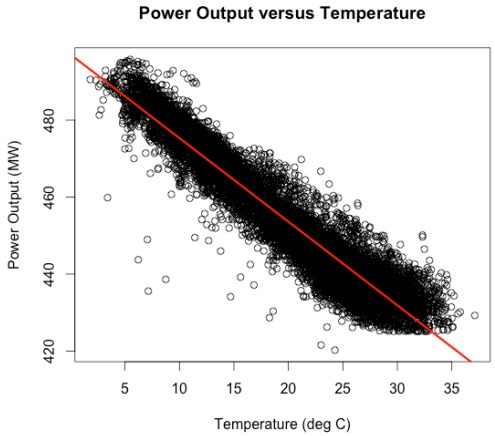


Figure 2. This graph shows the negative linear relationship between Power Output and Temperature, meaning that as Temperature increases, Power Output decreases. The red line is the linear regression line.

In order to diagnose the strength of the relationship, it is important to check whether the residuals (or errors) are normally distributed. Figure 3 shows three diagnosis plots. The QQ plot and the histogram both show the residuals are normally distributed. The Residuals versus Fitted Chart also shows that the residuals are close together, which is a sign of linearity.

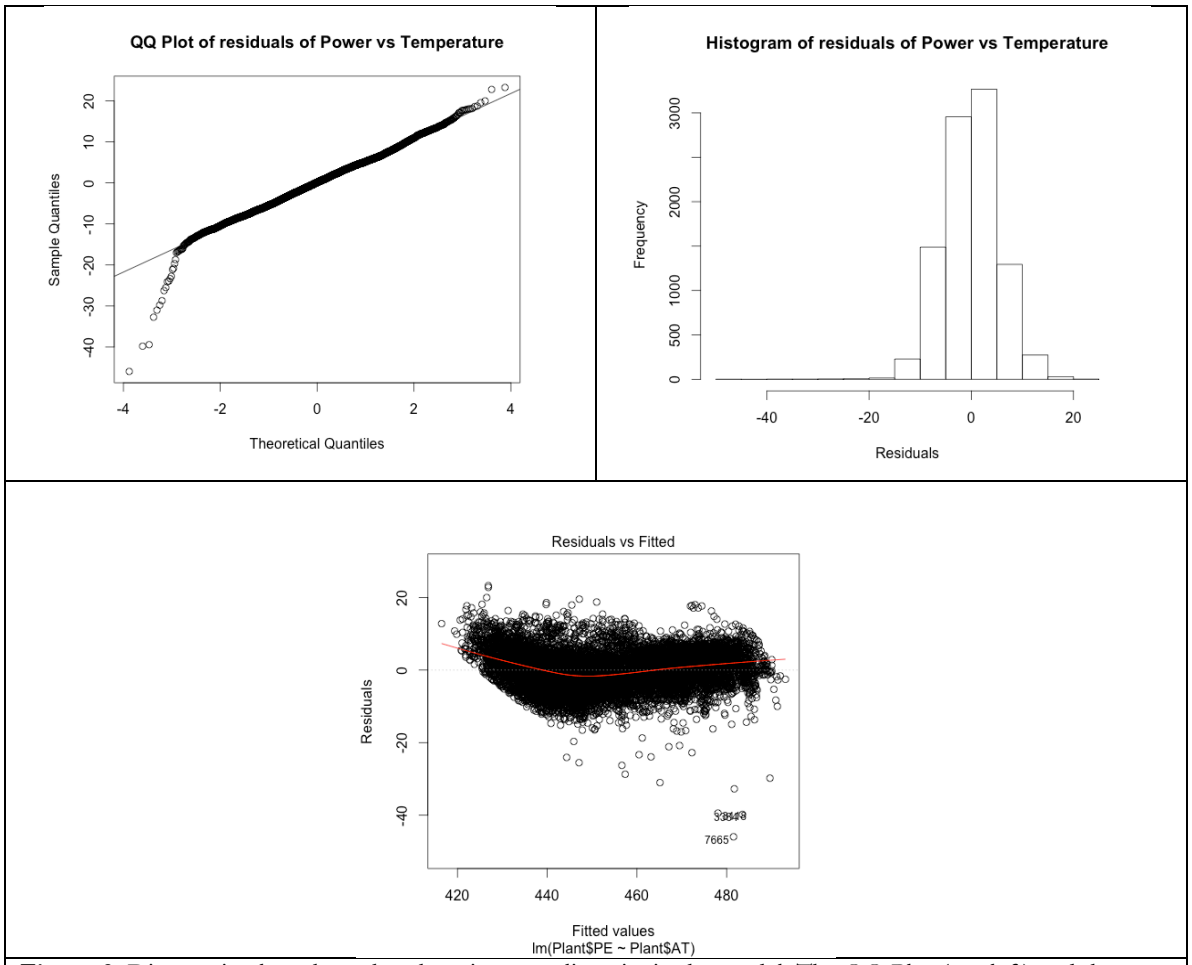


Figure 3. Diagnostic plots show that there is strong linearity in the model. The QQ Plot (top left) and the Histogram (top right) show that the residuals are normal. A plot of the Residuals versus Fitted values are clumped together, which shows linearity.

Relative Humidity effects on Power Output

The same analysis as performed to test the relationship between Relative Humidity and Power Output. This relationship is expressed as:

Power Output ~ *Relative Humidity* [2]

In this case, the linear regression model returned a low R^2 value of 0.1519, which means that Relative Humidity does not explain Power Output very well. Figure 4 shows a plot of the data with the regression line imposed on top.

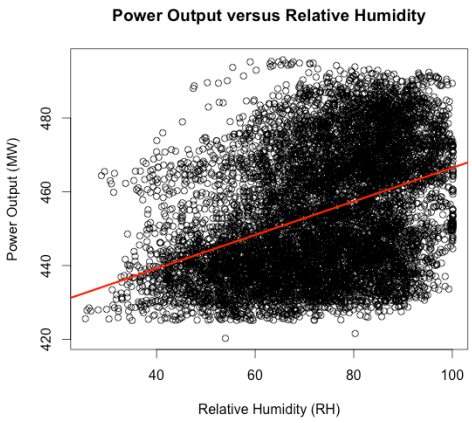
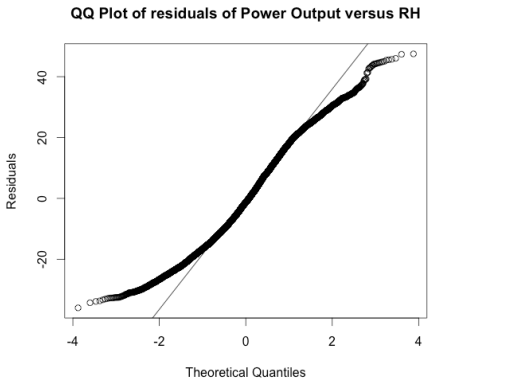
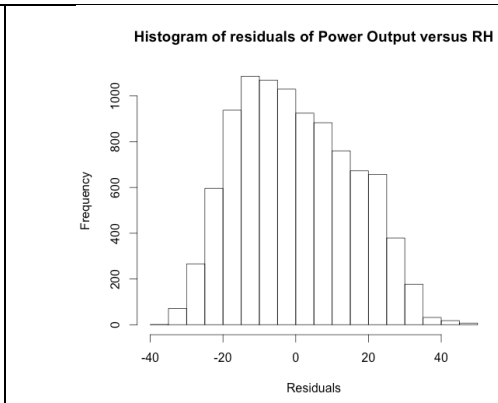


Figure 4. Power Output versus Relative Humidity plot with the linear regression imposed on top of the chart

Diagnostic plots were also created for this linear model. Note that the QQ plot is the greatest visual indicator of non-linearity, while the histogram and sequence plot are less obvious.





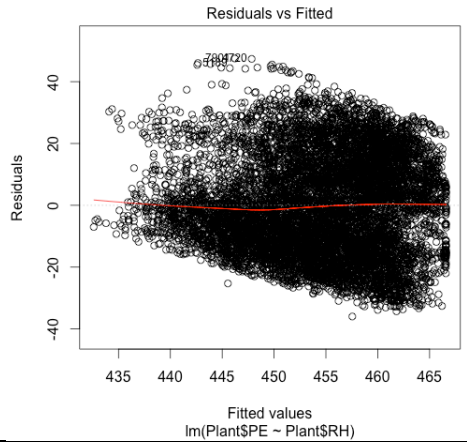


Figure 5. Diagnostic plots show that the residuals of the regression of Power Output versus Relative Humidity are not normal

Other factors which impact Power Output

Finally, all possible combinations were analyzed in a multiple regression model. The results showed that a model in which all parameters are considered explains 92.87% of Power Output, but the hypothesis test for Ambient Pressure is higher than all other variables. This indicates that Ambient Pressure can be removed from the model. A model with just Temperature, Relative Humidity, and Exhaust Volume also shows a very similar Adjusted R^2. Table 1 shows the R2 values that were returned for all combinations.

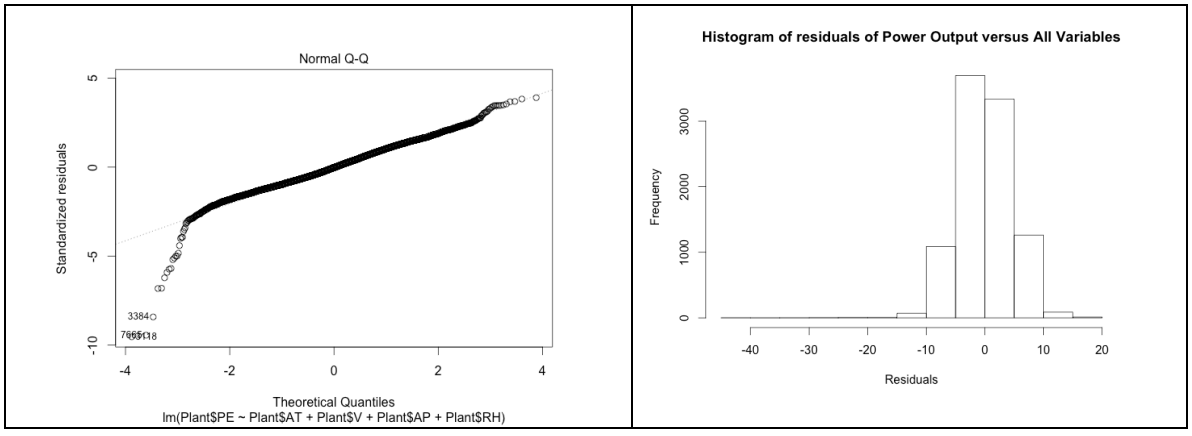
	Power Output ~	**Adjusted R^2**
1.	AT + AP + RH + V	0.9287
2.	T + RH + V	0.9284
3.	AT + V	0.9157
4.	AT + RH	0.9209
5.	V + RH	0.772

Table 1. R^2 generated from multiple linear regression models created for all possible combinations of variables

At this point, the AIC method was applied in both directions to decide between model 1 and model 2. After applying AIC, it was revealed that there is no difference between the model 1 and 2, so model 1 is selected for completeness:

$$Power\ Output \sim Temperature + Ambient\ Pressure + Relative\ Humidity + Exhaust\ Volume \quad [3]$$

Figure 6 shows the diagnosis plots that were created to test the model. Note the horizontal line in the Fitted vs. Residual graph shows linearity, the QQ plot is also a good linear fit, and that the histogram of the residuals is normal.



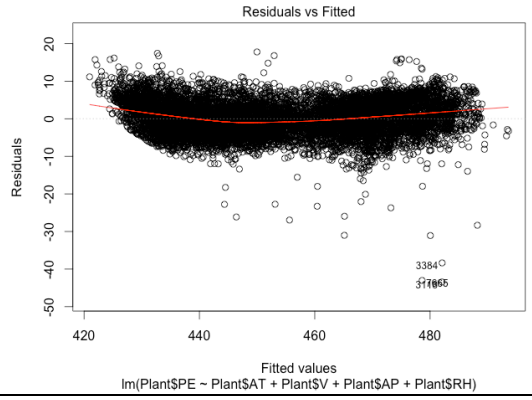


Figure 6. Diagnostic plots of the final model. The QQ Plot and the Histogram show a normal distribution of residuals. The plot of the Residuals versus Fitted values is clumped together, which shows a linear model

PART 2

Introduction

The following dataset comes from a researcher who has measured a number of rock samples, and is looking to understand rock strength. Figure 1 provides a sample of the data collected. Note that this dataset only contains 30 points, the impacts of which will be discussed further.

	ucs	qtz	plag	k.fds	hb	gs	ga	sf	ar
1	100.6	40.3	9.98	17.01	21.57	0.031	754.4	0.594	0.630
2	112.0	47.1	8.50	15.00	23.00	0.025	490.6	0.612	0.612

Figure 1. Snapshot of the dataset, which Uniaxial compressive strength (ucs), Quartz Content in % (qtz), Plagioclase Content in % (plag), Feldspar Content in %(k.fds), Hornblende Content in % (hb), Grain size in pixels/mm (gs), Grain Area in pixel/mm^2 (ga), Shape factor (sf), Aspect Ratio (ar)

Exploratory data analysis and multiple linear regression is performed on this dataset to find a model which best explains Uniaxial Compressive Strength.

Exploratory Data Analysis

To begin the analysis, a pairwise plot of all the data was created. As highlighted in Figure 2, linear relationships seem to exist between 'ucs' and 'qtz', 'hb', 'gs', 'ga', and 'sf'. This helps to develop an initial hypothesis that the model might involve these factors.

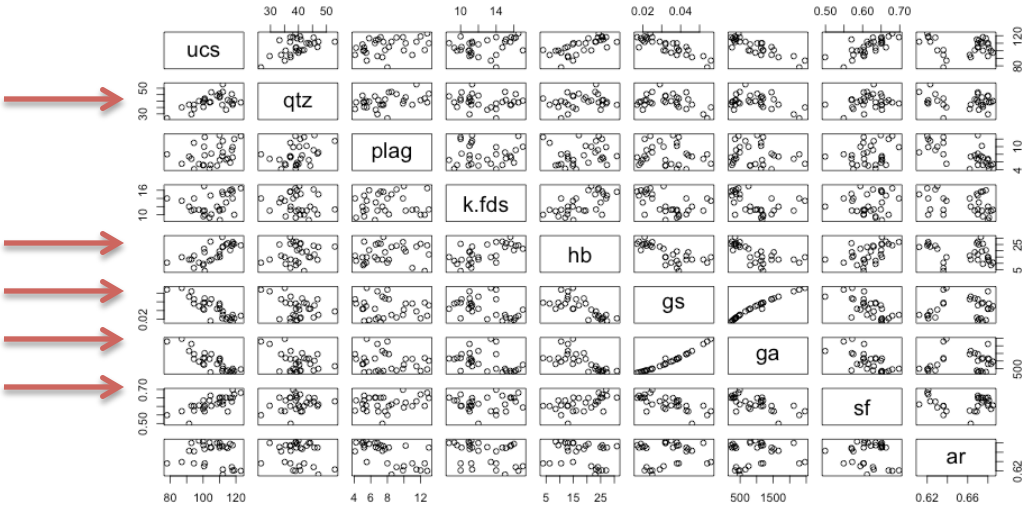


Figure 2. A pairwise plot shows that there are somewhat linear relationships between 'ucs' and 'hb', 'gs', 'ga', and 'sf'.

Multiple Regression Analysis

In order to understand whether the variables should be transformed, the Box-Cox method was run on the dataset. The Box-Cox method shows that lamba is -0.2. This value is close to zero, which signals that the data could be transformed to the log scale (when lambda equals zero, it is best to transform). However, since this dataset only includes 30 values, it was decided to not transform the data.

A multiple linear regression was performed on the model, including all the variables. This analysis only showed that quartz content and shape factor are significant for a model. Reviewing the pairs plot showed that other linear relationships exist within the data, so it was determined that another method to finding the model should be used.

Next, the Akaike Information Criteria (AIC) method was performed in the forward and backwards direction. The results of the AIC test state the model should be the following:

$$Rocks\$UCS \sim Rocks\$qtz + Rocks\$plag + Rocks\$hb + Rocks\$gs + Rocks\$sf + Rocks\$ar \quad [1]$$

However, the summary of the AIC showed that the Plagioclase Content and Aspect Ratio are not significant. The AIC method was run two more times: once only removing Aspect Ratio and then with only removing Plagioclase Content in equation 1. Both times the models ran, the results showed that Hornblende Content is insignificant. Therefore, the final model that explains Uniaxial Compressive Strength is:

$$Uniaxial\ Compressive\ Strength \sim Quartz\ Content + Grain\ Size + Shape\ Factor \quad [2]$$

This model was confirmed through inspecting the diagnostic charts given in Figure 3.

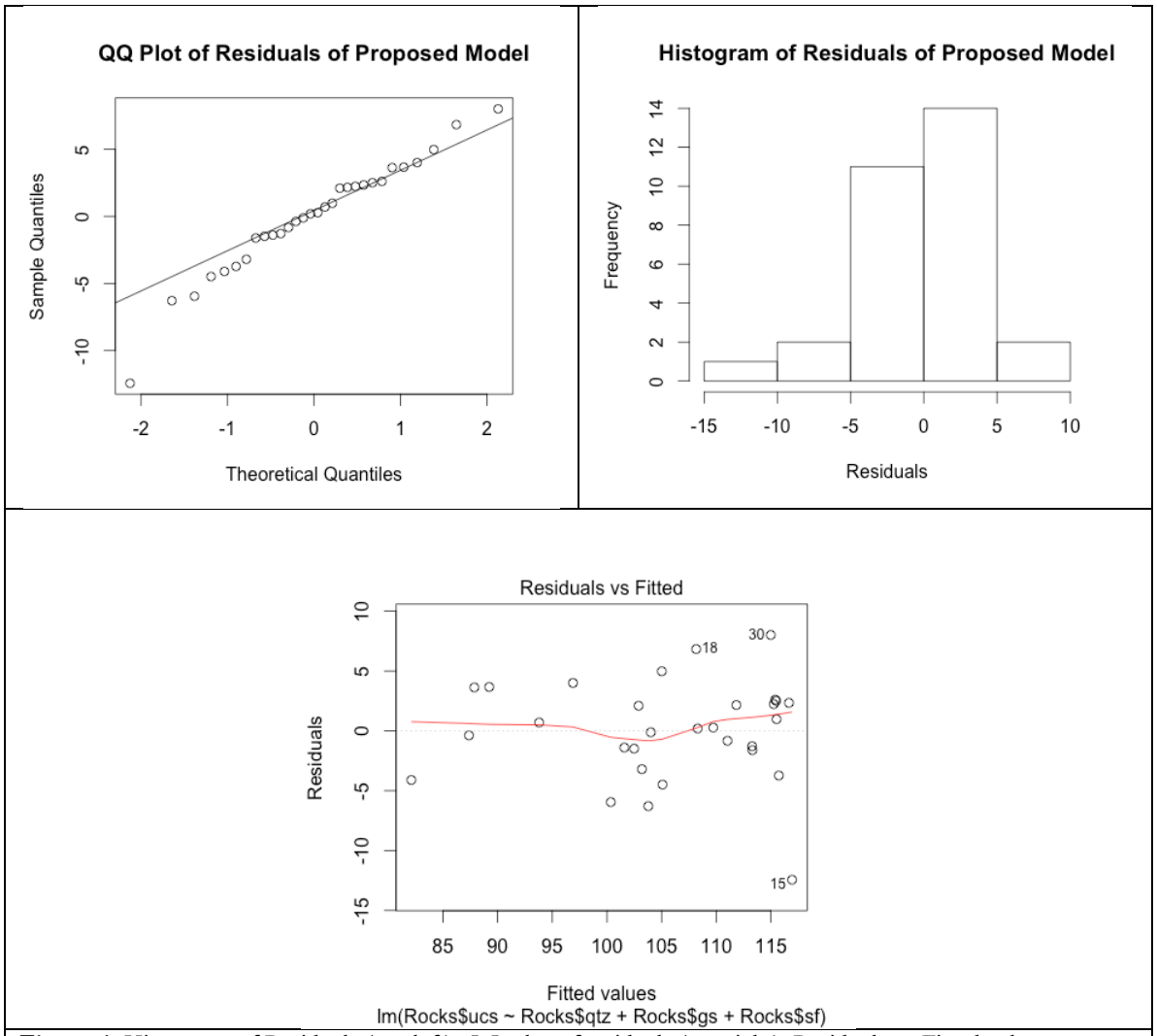


Figure 4. Histogram of Residuals (top left), QQ plot of residuals (top right), Residuals vs Fitted values (bottom center)

The QQ plot of the residuals falls into a linear line, which represents the mostly normal distribution in the histogram of residuals. The Residuals versus Fitted Values chart also shows that the residuals are mostly clumped together, which also indicates a linear fit for the model.

These results were compared to QQ plots for other possible combinations of model selection. Note that the Proposed Model has the best linear fit out of all the possibilities.

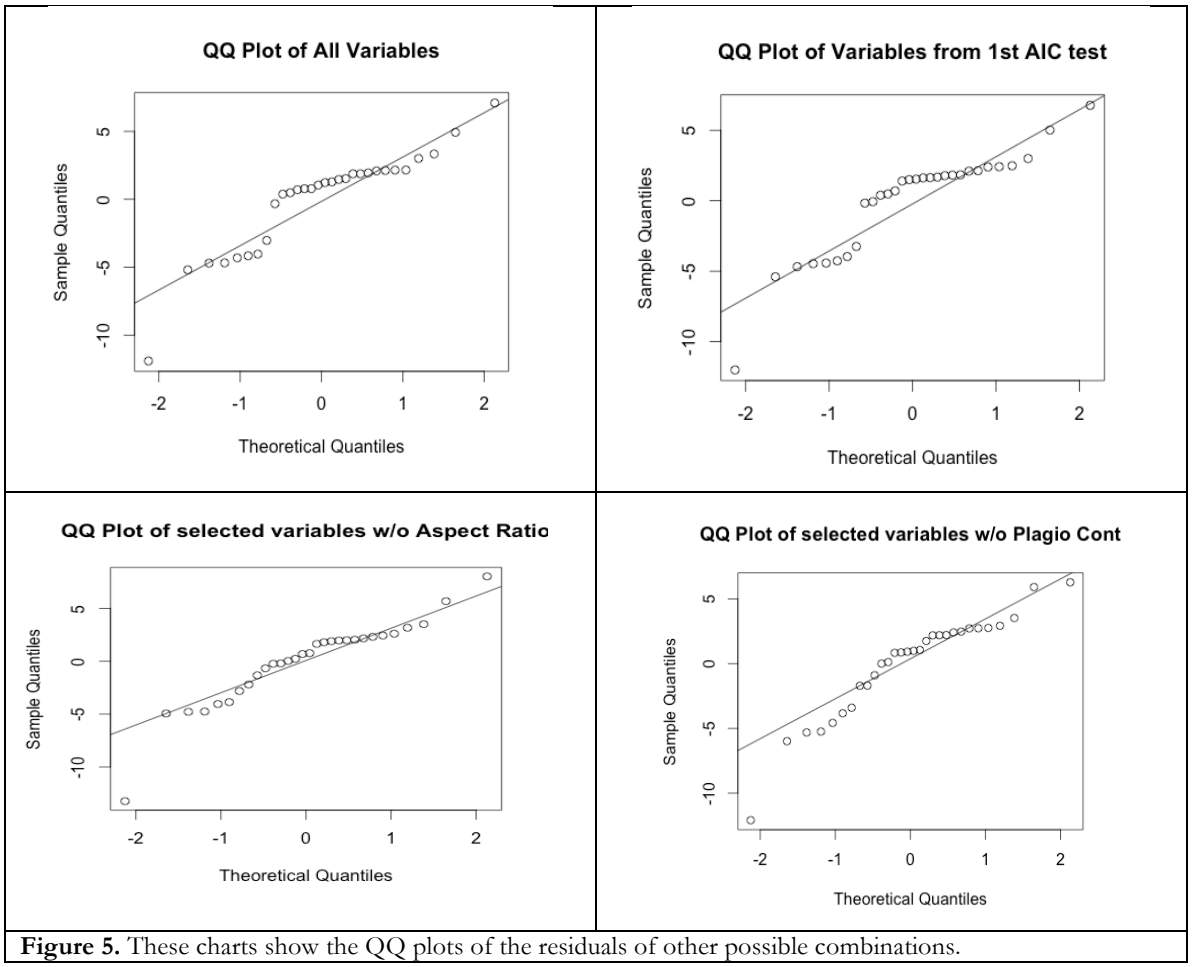


Figure 5. These charts show the QQ plots of the residuals of other possible combinations.

Limitations

The greatest limitation in this model is the fact that there is very limited data. Although the AIC method was used to find the model, there isn't enough data to truly confirm this relationship. Furthermore, the histogram and QQ plot of the final model are more normal than compared to other models. However, more data could change this analysis. Therefore this analysis provides a strong hypothesis for what the model could be, and with more data this hypothesis can be tested further.

R Code

PART 1

Figure 2

```
#Linear Regression
powerAll.lm <- lm(Plant$PE ~ Plant$AT + Plant$V + Plant$AP + Plant$RH)
summary(powerAll.lm)
hist(powerAll.lm$residuals, main = "Histogram of residuals of Power
Output versus All Variables", xlab = "Residuals")pairs(Rocks)
```

Figure 3

```
#Residual error analysis - sequence plot with residual
plot(residuals(powerTemp.lm), main = "Sequence plot of residuals of
Power vs Temperature", ylab = "Residuals")
plot(powerTemp.lm$fitted.values, powerTemp.lm$residuals, main = "Fitted
Values vs. Residuals", xlab = "Fitted Values", ylab = "Residuals")

#QQ Plot of residuals
qqnorm(residuals(powerTemp.lm), main = "QQ Plot of residuals of Power
vs Temperature")
qqline(residuals(powerTemp.lm))

#Histogram of resduals
hist(residuals(powerTemp.lm), main = "Histogram of residuals of Power
vs Temperature", xlab = "Residuals")

#Linear Regression diagnostic
plot(powerTemp.lm)
```

Figure 4

```
#Linear Regression
powerHum.lm <- lm(Plant$PE ~ Plant$RH)
summary(powerHum.lm)
plot(Plant$RH, Plant$PE, main = "Power Output versus Relative
Humidity", xlab = "Relative Humidity (RH)", ylab = "Power Output (MW)")
abline(powerHum.lm, col = 2, lwd = 3)
```

Figure 5

```
#Residual error analysis - sequence plot with residual
plot(residuals(powerHum.lm), main = "Sequence plot of residuals of
Power Output versus RH")

#QQ Plot of residuals
qqnorm(residuals(powerHum.lm), main = "QQ Plot of residuals of Power
Output versus RH", ylab = "Residuals")
qqline(residuals(powerHum.lm))

#Histogram of resduals
hist(residuals(powerHum.lm), main = "Histogram of residuals of Power
Output versus RH", xlab = "Residuals")

#Linear regression diagnostic plots
plot(powerHum.lm)
#Linear Regression
```

Figure 6

```
powerAll.lm <- lm(Plant$PE ~ Plant$AT + Plant$V + Plant$AP + Plant$RH)
summary(powerAll.lm)
hist(powerAll.lm$residuals, main = "Histogram of residuals of Power
Output versus All Variables", xlab = "Residuals")

powerATVRH.lm <- lm(Plant$PE ~ Plant$AT + Plant$V + Plant$RH)
summary(powerATVRH.lm)
plot(residuals(powerATVRH.lm))
qqnorm(residuals(powerATVRH.lm))
hist(residuals(powerATVRH.lm))
#R2 = 0.9284, basically stays the same with AP removed, Residuals plot
looks pretty good
#QQ and hist also seem gaussian
#Check the difference between having AP, but this the best one
#Not a lot of significance because the numbers are so small

#Use AIC to determine model
powerAllAIC <- stepAIC(powerAll.lm)
summary(powerAllAIC)
plot(powerAllAIC)
#shows that AP could be removed

#Use AIC to determine model without AP
powerATVRH.AIC.lm <- stepAIC(powerATVRH.lm, direction = "both")
summary(powerATVRH.AIC.lm)
plot(powerATVRH.AIC.lm)

#Check other models
powerATV.lm <- lm(Plant$PE ~ Plant$AT + Plant$V)
summary(powerATV.lm)
plot(residuals(powerATV.lm))
#R2 = 0.9157, lower without RH
#Residuals plot looks the same as ATVRH

#Check other models
powerATRH.lm <- lm(Plant$PE ~ Plant$AT + Plant$RH)
summary(powerATRH.lm)
plot(residuals(powerATRH.lm))
#R2 = 0.9209
#Residuals plot looks the same as ATVRH

#Check other models
powerVRH.lm <- lm(Plant$PE ~ Plant$V + Plant$RH)
summary(powerVRH.lm)
plot(residuals(powerATRH.lm))
#R2 = 0.772, very low
#Residuals plot looks okay
```

PART 2

Figure 2

```
pairs(Rocks)
```

Multiple Linear Regression Analysis

```
#Find out whether Rocks$ucs needs to be transformed
boxcox(ucs~1, data = Rocks, lambda = seq(-2,  2, 1/10))
#Lamba = -0.2, therefore we should not transform to log base

#Rock Strength Linear Regression
ucs.lm <- lm(Rocks$ucs ~ Rocks$qtz + Rocks$plag + Rocks$k.fds +
Rocks$hb + Rocks$gs + Rocks$ga + Rocks$sf + Rocks$ar)
summary(ucs.lm)
plot(residuals(ucs.lm))
qqnorm(residuals(ucs.lm))
hist(residuals(ucs.lm))
#R2 = 0.8296

#Run AIC on non logform because the lm wasn't providing a sensible
conclusion
stepratio <- stepAIC(ucs.lm, direction = "both")
summary(stepratio)
qqnorm(ucs.lm$residuals, main = "QQ Plot of All Variables")
qqline(ucs.lm$residuals)

#AIC returns qtz, plag, hb, gs, sf, ar
ucsAIC.lm <- lm(Rocks$ucs ~ Rocks$qtz + Rocks$plag + Rocks$hb +
Rocks$gs + Rocks$sf + Rocks$ar)
qqnorm(ucsAIC.lm$residuals, main = "QQ Plot of Variables from 1st AIC
test")
qqline(ucsAIC.lm$residuals)

#Run AIC on stepratio
stepratio2 <- stepAIC(ucsAIC.lm, direction = "both")
summary(stepratio2)

#stepratio1 says that plag and ar are not significant. Start by
removing ar and see what happens
ucsAIC1.lm <- lm(Rocks$ucs ~ Rocks$qtz + Rocks$plag + Rocks$hb +
Rocks$gs + Rocks$sf)
stepratio3 <- stepAIC(ucsAIC1.lm, direction = "both")
summary(stepratio3)
#stepratio3 says that the model should not include hb.

#let's see what happens if i keep plag, and remove ar
ucsAIC2.lm <- lm(Rocks$ucs ~ Rocks$qtz + Rocks$hb + Rocks$gs + Rocks$sf
+ Rocks$ar)
stepratio4 <- stepAIC(ucsAIC2.lm, direction = "both")
summary(stepratio4)
#stepratio4 returns qtz, hb, gs, sf
hist(ucsAIC2.lm$residuals)
qqnorm(ucsAIC2.lm$residuals, main = "QQ Plot of selected variables w/o
Aspect Ratio")
qqline(ucsAIC2.lm$residuals)
#summary(stepratio4) says to keep qtz, gs, sf
```

```
#lets see what happens if i keep ar and remove plag
ucsAIC3.lm <- lm(Rocks$ucs ~ Rocks$qtz + Rocks$hb + Rocks$gs + Rocks$sf
+ Rocks$plag)
stepratio5 <- stepAIC(ucsAIC3.lm, direction = "both")
summary(stepratio5)
#summary(stepratio5) says to keep qtz, gs, sf --- so hb is out
qqnorm(ucsAIC3.lm$residuals, main = "QQ Plot of selected variables w/o
Plagio Cont")
qqline(ucsAIC3.lm$residuals)

#lets run an lm on qtz, gs, sf
ucsAIC4.lm <- lm(Rocks$ucs ~ Rocks$qtz + Rocks$gs + Rocks$sf)
stepratio6 <- stepAIC(ucsAIC4.lm)
summary(stepratio6)
#stepratio5 returns
hist(ucsAIC4.lm$residuals, main = "Histogram of Residuals of Proposed
Model", xlab  = "Residuals")
qqnorm(ucsAIC4.lm$residuals, main = "QQ Plot of Residuals of Proposed
Model")
qqline(ucsAIC4.lm$residuals)
```