Gas Turbine CO and NOx Emission Analysis

Aayushi Gupta, Kyle Kaminski, Rosa Lin, Ruben Martinez

Introduction

Gas Turbine CO and NOx Emission Data Set

The data comes from a gas turbine located in Turkey that studies the flue gas emissions of specifically carbon monoxide (CO) and nitrogen oxide (NOx) gases. The data set provides hourly statistics of 11 sensors. Data points were collected from a gas turbine from Jan 01 2011 to Dec 13 2015.

Description

The data file `gt_2015.csv` has 7384 observations and 11 variables from the UCI Gas Turbine CO and NOx Emission Data Set. We are going to explore and analyze the following variables:

- `AT` - Ambient Temperature
- `AP` - Ambient Pressure
- `AH` - Ambient Humidity
- `AFDP` - Air filter difference pressure
- `GTEP` - Gas turbine exhaust pressure
- `TIT` - Turbine inlet temperature
- `TAT` - Turbine after temperature
- `TEY` - Turbine energy yield
- `CDP` - Compressor discharge pressure

Here's a quick peek at the data set:

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
1.95320	1020.1	84.985	2.5304	20.116	1048.7	544.92	116.27	10.799	7.4491	113.250
1.21910	1020.1	87.523	2.3937	18.584	1045.5	548.50	109.18	10.347	6.4684	112.020
0.94915	1022.2	78.335	2.7789	22.264	1068.8	549.95	125.88	11.256	3.6335	88.147
1.00750	1021.7	76.942	2.8170	23.358	1075.2	549.63	132.21	11.702	3.1972	87.078
1.28580	1021.6	76.732	2.8377	23.483	1076.2	549.68	133.58	11.737	2.3833	82.515
1.83190	1021.7	76.411	2.8410	23.495	1076.4	549.92	133.58	11.829	2.0812	81.193

Goals

The goal for this project is to utilize this data set for the purpose of studying flue gas emissions, specifically carbon monoxide(CO) and nitrogen oxides (NOx). Our focus will be to find statistically significant relationships between the ambient and turbine variables and the emissions variables. We will limit the size of our model to more clearly demonstrate these relationships. Ultimately we will suggest which variables make the biggest impact on emission levels in order to decrease emissions overall.

#Exploratory Data Analysis

Relationships between feature variables

Figure 1: Scatterplot Matrices to decide which feature variables have a linear relationship

```
pairs(gt_2015, pch = 20, cex = 0.25)
```

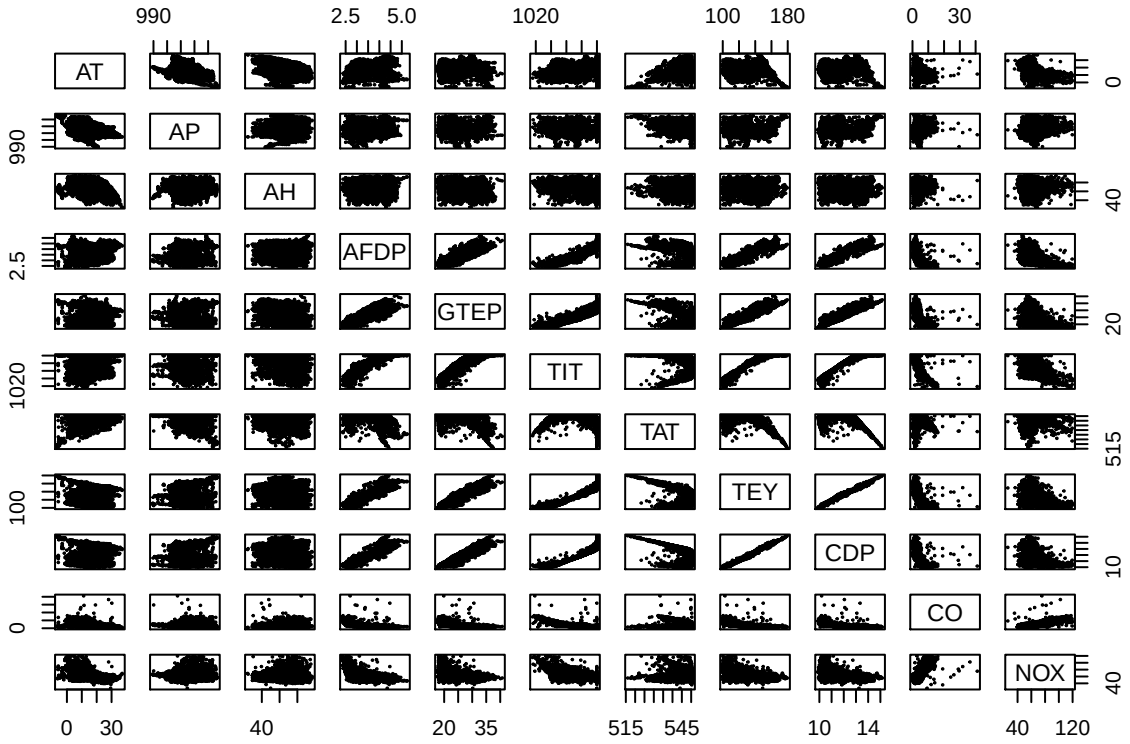


Figure 2:

```
knitr::kable(cor(gt_2015), digits = 2, caption = "Pairwise Correlation Between Variables")
```

Table 2: Pairwise Correlation Between Variables

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
AT	1.00	-0.49	-0.47	0.47	0.19	0.33	0.21	0.11	0.20	-0.39	-0.59
AP	-0.49	1.00	0.08	-0.09	-0.04	-0.08	-0.29	0.05	0.03	0.20	0.21
AH	-0.47	0.08	1.00	-0.25	-0.30	-0.26	0.03	-0.18	-0.22	0.16	0.07
AFDP	0.47	-0.09	-0.25	1.00	0.84	0.92	-0.52	0.88	0.92	-0.64	-0.58
GTEP	0.19	-0.04	-0.30	0.84	1.00	0.89	-0.62	0.93	0.94	-0.56	-0.37
TIT	0.33	-0.08	-0.26	0.92	0.89	1.00	-0.40	0.95	0.95	-0.74	-0.52
TAT	0.21	-0.29	0.03	-0.52	-0.62	-0.40	1.00	-0.63	-0.66	0.03	0.05
TEY	0.11	0.05	-0.18	0.88	0.93	0.95	-0.63	1.00	0.99	-0.62	-0.40
CDP	0.20	0.03	-0.22	0.92	0.94	0.95	-0.66	0.99	1.00	-0.61	-0.44
CO	-0.39	0.20	0.16	-0.64	-0.56	-0.74	0.03	-0.62	-0.61	1.00	0.68
NOX	-0.59	0.21	0.07	-0.58	-0.37	-0.52	0.05	-0.40	-0.44	0.68	1.00

Remove variables that are highly correlated.

```
#vif
library(faraway)
model = lm(CO ~ . - NOX - TIT - CDP - TEY, data = gt_2015)
vif(model)
```

```
##       AT       AP       AH     AFDP     GTEP      TAT
## 3.866424 1.600597 1.718769 7.412520 5.909197 2.301015
```

Methods

Linear Regression

We will create a multiple linear regression model using all feature variables mentioned in the description of Section 1. The implementation and parameters of this model can be obtained by the following equation where we will find estimates for the parameters

$$\beta$$

using:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Key assumptions are stated as:

- **L**inearity: can be written as a linear combination of the predictors.
- **I**ndependence: the errors are independent of each other (not highly correlated).
- **N**ormality: the distribution of the errors follow a normal distribution.
- **E**qual Variance: the error variance is the same.[1]

We will then use model selection using backward BIC to tune our model and remove any insignificant predictor variables. This selection prefers smaller models which aligns with our goal of limiting the size of our final model.

```
full_model = lm(CO ~ ., data = gt_2015)
linear_model = lm(CO ~ .-NOX - TIT - CDP - TEY, data = gt_2015)
summary(linear_model)
```

```
##
## Call:
## lm(formula = CO ~ . - NOX - TIT - CDP - TEY, data = gt_2015)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.839 -0.673 -0.132  0.481 34.242
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 145.099775   4.408695  32.912  < 2e-16 ***
```

[1] Dalpiaz David, Applied Statistics in R, https://daviddalpiaz.github.io/appliedstats/model-diagnostics.html

```
## AT            0.028276   0.004060   6.965 3.57e-12 ***
## AP            0.001918   0.003067   0.625    0.532
## AH           -0.009753   0.001618  -6.026 1.76e-09 ***
## AFDP         -2.531044   0.074576 -33.939  < 2e-16 ***
## GTEP         -0.186308   0.009082 -20.513  < 2e-16 ***
## TAT          -0.237369   0.004619 -51.387  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.436 on 7377 degrees of freedom
## Multiple R-squared:  0.5874, Adjusted R-squared:  0.587
## F-statistic:  1750 on 6 and 7377 DF,  p-value: < 2.2e-16
```

```
#picking a new variable to test
AT_model = lm(CO ~ AT, data = gt_2015)
AP_model = lm(CO ~ AP, data = gt_2015)
AH_model = lm(CO ~ AH, data = gt_2015)
AFDP_model = lm(CO ~ AFDP, data = gt_2015)
GTEP_model = lm(CO ~ GTEP, data = gt_2015)
TAT_model = lm(CO ~ TAT, data = gt_2015)
BIC(AT_model)
```

```
## [1] 31634.69
```

```
BIC(AP_model)
```

```
## [1] 32553.05
```

```
BIC(AH_model)
```

```
## [1] 32668.32
```

```
BIC(AFDP_model) #second best
```

```
## [1] 28953.71
```

```
BIC(GTEP_model)
```

```
## [1] 30112.68
```

```
BIC(TAT_model)
```

```
## [1] 32852.49
```

```
BIC(linear_model) #this is the best model
```

```
## [1] 26365.45
```

```r
library(MASS)

n = length(resid(linear_model))
BIC_model = step(linear_model, direction = "backward", k = log(n))
```

```
## Start:  AIC=5401.66
## CO ~ (AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP + NOX) -
##     NOX - TIT - CDP - TEY
##
##        Df Sum of Sq   RSS    AIC
## - AP    1       0.8 15218 5393.1
## <none>              15217 5401.7
## - AH    1      74.9 15292 5429.0
## - AT    1     100.1 15317 5441.1
## - GTEP  1     868.0 16085 5802.4
## - AFDP  1    2376.0 17593 6464.1
## - TAT   1    5447.0 20664 7652.1
##
## Step:  AIC=5393.14
## CO ~ AT + AH + AFDP + GTEP + TAT
##
##        Df Sum of Sq   RSS    AIC
## <none>              15218 5393.1
## - AH    1      86.8 15304 5426.2
## - AT    1     120.5 15338 5442.5
## - GTEP  1     991.0 16209 5850.1
## - AFDP  1    2564.3 17782 6534.1
## - TAT   1    5608.5 20826 7701.0
```

```r
coef(BIC_model)
```

```
##  (Intercept)           AT           AH         AFDP         GTEP          TAT
## 147.33755305   0.02702808  -0.01004686  -2.51758434  -0.18816259  -0.23782668
```

```r
stepAIC(linear_model, direction = "backward")
```

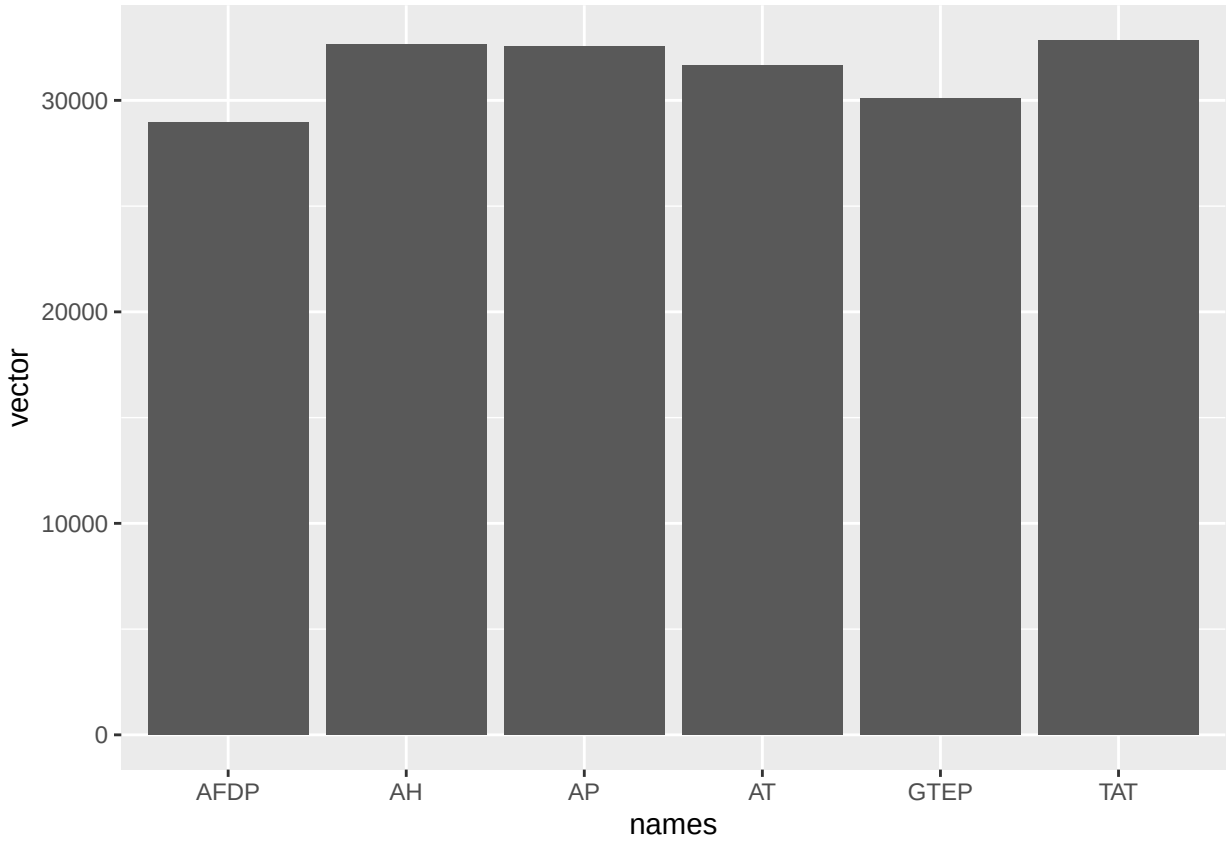
```
## Start:  AIC=5353.31
## CO ~ (AT + AP + AH + AFDP + GTEP + TIT + TAT + TEY + CDP + NOX) -
##     NOX - TIT - CDP - TEY
##
##        Df Sum of Sq   RSS    AIC
## - AP    1       0.8 15218 5351.7
## <none>              15217 5353.3
## - AH    1      74.9 15292 5387.6
## - AT    1     100.1 15317 5399.7
## - GTEP  1     868.0 16085 5760.9
## - AFDP  1    2376.0 17593 6422.6
## - TAT   1    5447.0 20664 7610.7
##
## Step:  AIC=5351.7
## CO ~ AT + AH + AFDP + GTEP + TAT
##
```

```
##        Df Sum of Sq   RSS    AIC
## <none>              15218 5351.7
## - AH    1      86.8 15304 5391.7
## - AT    1     120.5 15338 5408.0
## - GTEP  1     991.0 16209 5815.5
## - AFDP  1    2564.3 17782 6499.6
## - TAT   1    5608.5 20826 7666.5
```

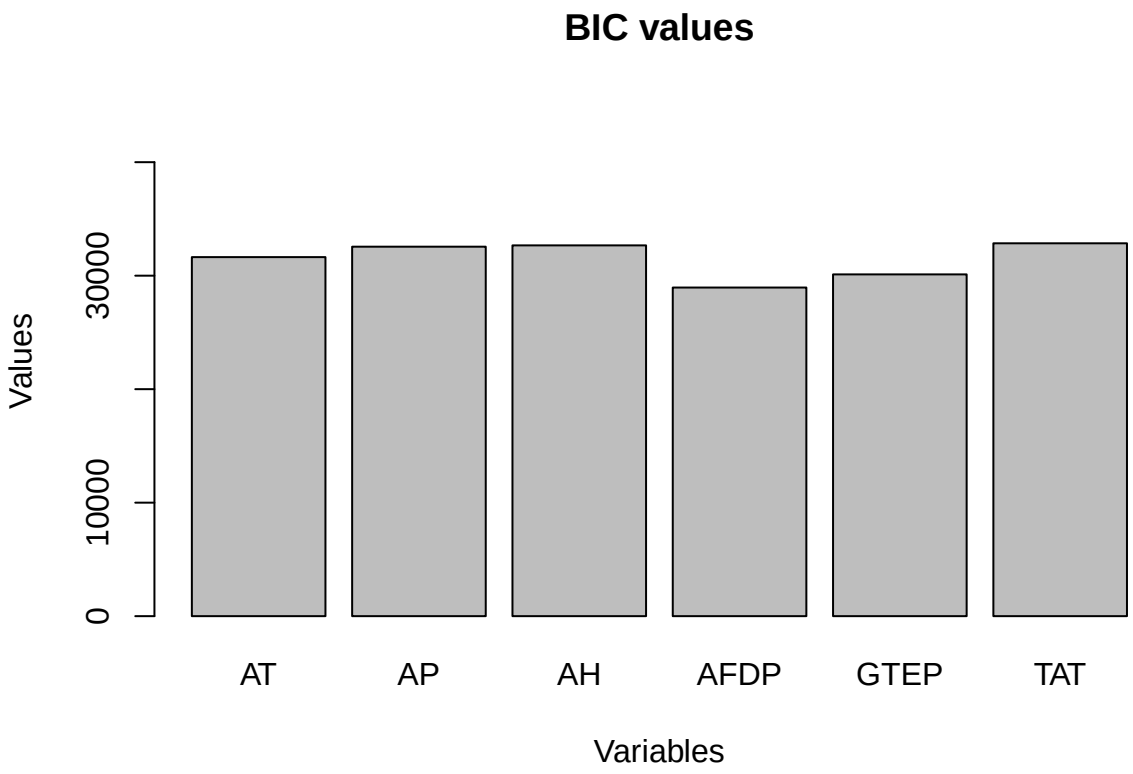
```
##
## Call:
## lm(formula = CO ~ AT + AH + AFDP + GTEP + TAT, data = gt_2015)
##
## Coefficients:
## (Intercept)           AT           AH         AFDP         GTEP          TAT
##   147.33755      0.02703     -0.01005     -2.51758     -0.18816     -0.23783
```

```
vector <- c(BIC(AT_model), BIC(AP_model), BIC(AH_model), BIC(AFDP_model), BIC(GTEP_model), BIC(TAT_model

library(ggplot2)
df <- data.frame(vector = c(BIC(AT_model), BIC(AP_model), BIC(AH_model), BIC(AFDP_model), BIC(GTEP_model
ggplot(data = df, aes(x = names, y = vector), ylim = c(0, 50000)) + geom_bar(stat = "identity")
```




```
barplot(vector, main = "BIC values", xlab = "Variables",ylab = "Values", names.arg = c("AT", "AP", "AH"
```


BIC values
Values
0
10000
30000
AT
AP
AH
AFDP
GTEP
TAT
Variables