Gas Turbine CO Emission Analysis

Aayushi Gupta, Kyle Kaminski, Rosa Lin, Ruben Martinez

Introduction

The combined cycle power plant, also known as combined cycle gas turbine plant, is an assembly of heat engines that combine to generate electricity (Tüfekci). A combined-cycle power plant (CCPP) is made up of gas turbines, steam turbines, and heat recovery steam generators. The electricity is generated and combined in one cycle by gas and steam turbines and then transferred from one turbine to another.

We are interested in identifying the process variables that impact carbon monoxide emissions. By determining the process variables that impact carbon monoxide emissions we will be able to find opportunities to reduce carbon monoxide emissions.

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

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
1.0075	1021.7	76.942	2.8170	23.358	1075.2	549.63	132.21	11.702	3.1972	87.078

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
1.2858	1021.6	76.732	2.8377	23.483	1076.2	549.68	133.58	11.737	2.3833	82.515
1.8319	1021.7	76.411	2.8410	23.495	1076.4	549.92	133.58	11.829	2.0812	81.193
2.0740	1022.0	75.974	2.7981	22.945	1073.7	549.98	131.53	11.687	2.2529	83.171
1.7824	1022.6	73.535	2.8327	23.337	1075.7	550.01	133.18	11.745	3.7350	85.749
1.7797	1025.1	68.528	2.8725	23.276	1077.0	550.03	134.21	11.782	3.6902	85.317

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
0.68481	1026.7	56.029	4.0703	34.213	1100.0	527.71	168.83	14.358	2.9790	59.354
1.99950	1026.3	54.000	3.9830	34.171	1100.1	530.64	166.13	14.182	2.5793	59.432
3.16630	1025.7	51.350	4.0683	35.162	1099.8	528.21	167.49	14.384	2.2228	58.432
3.64860	1025.3	51.649	4.0375	35.282	1100.0	530.04	165.89	14.257	2.2119	60.172
3.92070	1025.2	49.619	4.0455	34.648	1099.9	529.73	166.00	14.253	2.3487	60.255
3.91930	1025.1	51.181	4.0400	33.944	1100.0	529.56	166.46	14.283	2.3095	59.778

Here's some descriptive statistics of the data set:

```
##        AT                AP                AH              AFDP
##  Min.   :-6.235   Min.   : 989.4   Min.   :24.09   Min.   :2.369
##  1st Qu.:11.073   1st Qu.:1009.7   1st Qu.:59.45   1st Qu.:3.117
##  Median :17.456   Median :1014.0   Median :70.95   Median :3.538
##  Mean   :17.225   Mean   :1014.5   Mean   :68.65   Mean   :3.599
##  3rd Qu.:23.685   3rd Qu.:1018.3   3rd Qu.:79.65   3rd Qu.:4.195
##  Max.   :37.103   Max.   :1036.6   Max.   :96.67   Max.   :5.239
##       GTEP             TIT             TAT             TEY
##  Min.   :17.70   Min.   :1016   Min.   :516.0   Min.   :100.0
##  1st Qu.:23.15   1st Qu.:1070   1st Qu.:544.7   1st Qu.:126.3
##  Median :25.33   Median :1080   Median :549.7   Median :131.6
##  Mean   :26.13   Mean   :1079   Mean   :546.6   Mean   :134.0
##  3rd Qu.:30.02   3rd Qu.:1100   3rd Qu.:550.0   3rd Qu.:147.2
##  Max.   :40.72   Max.   :1100   Max.   :550.6   Max.   :179.5
##       CDP               CO               NOX
##  Min.   : 9.871   Min.   : 0.2128   Min.   : 25.91
##  1st Qu.:11.466   1st Qu.: 1.8082   1st Qu.: 52.40
##  Median :11.933   Median : 2.5334   Median : 56.84
##  Mean   :12.097   Mean   : 3.1300   Mean   : 59.89
##  3rd Qu.:13.148   3rd Qu.: 3.7026   3rd Qu.: 65.09
##  Max.   :15.159   Max.   :41.0970   Max.   :119.68
```

```
##        AT                AP                AH              AFDP
##  Min.   :-5.82   Min.   : 990.8   Min.   :31.62   Min.   :2.644
##  1st Qu.:10.51   1st Qu.:1010.6   1st Qu.:62.93   1st Qu.:3.207
##  Median :15.38   Median :1015.1   Median :72.87   Median :3.350
##  Mean   :15.57   Mean   :1015.1   Mean   :71.16   Mean   :3.436
##  3rd Qu.:20.73   3rd Qu.:1019.1   3rd Qu.:80.71   3rd Qu.:3.736
##  Max.   :35.41   Max.   :1035.6   Max.   :96.67   Max.   :4.291
##       GTEP             TIT             TAT             TEY              CDP
##  Min.   :21.70   Min.   :1052   Min.   :529.2   Min.   :127.0   Min.   :11.29
##  1st Qu.:23.63   1st Qu.:1075   1st Qu.:549.8   1st Qu.:129.8   1st Qu.:11.66
##  Median :24.26   Median :1077   Median :550.0   Median :130.5   Median :11.77
##  Mean   :24.66   Mean   :1077   Mean   :549.9   Mean   :130.6   Mean   :11.78
##  3rd Qu.:25.11   3rd Qu.:1080   3rd Qu.:550.1   3rd Qu.:131.3   3rd Qu.:11.90
```

```
##  Max.   :32.46   Max.   :1088   Max.   :550.5   Max.   :134.5   Max.   :12.38
##        CO               NOX
##  Min.   : 0.2573   Min.   : 35.60
##  1st Qu.: 2.0296   1st Qu.: 52.79
##  Median : 2.6383   Median : 57.88
##  Mean   : 2.7589   Mean   : 59.90
##  3rd Qu.: 3.4307   3rd Qu.: 66.77
##  Max.   :36.4540   Max.   :102.33
```

```
##        AT               AP             AH             AFDP
##  Min.   :-6.235   Min.   :1006   Min.   :44.25   Min.   :3.608
##  1st Qu.: 3.041   1st Qu.:1021   1st Qu.:67.35   1st Qu.:4.084
##  Median : 5.996   Median :1024   Median :75.36   Median :4.201
##  Mean   : 5.277   Mean   :1023   Mean   :74.19   Mean   :4.293
##  3rd Qu.: 8.255   3rd Qu.:1025   3rd Qu.:82.20   3rd Qu.:4.516
##  Max.   :15.830   Max.   :1037   Max.   :93.52   Max.   :5.239
##       GTEP            TIT            TAT             TEY             CDP
##  Min.   :30.99   Min.   :1096   Min.   :516.0   Min.   :160.0   Min.   :13.63
##  1st Qu.:32.81   1st Qu.:1100   1st Qu.:529.7   1st Qu.:162.0   1st Qu.:14.00
##  Median :33.43   Median :1100   Median :531.6   Median :163.8   Median :14.13
##  Mean   :33.72   Mean   :1100   Mean   :530.9   Mean   :164.7   Mean   :14.17
##  3rd Qu.:34.21   3rd Qu.:1100   3rd Qu.:533.3   3rd Qu.:166.4   3rd Qu.:14.25
##  Max.   :40.72   Max.   :1100   Max.   :538.5   Max.   :179.5   Max.   :15.16
##        CO              NOX
##  Min.   :0.2128   Min.   :45.92
##  1st Qu.:2.1237   1st Qu.:50.27
##  Median :2.4638   Median :57.37
##  Mean   :2.3864   Mean   :55.55
##  3rd Qu.:2.7363   3rd Qu.:59.70
##  Max.   :4.0948   Max.   :69.88
```

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

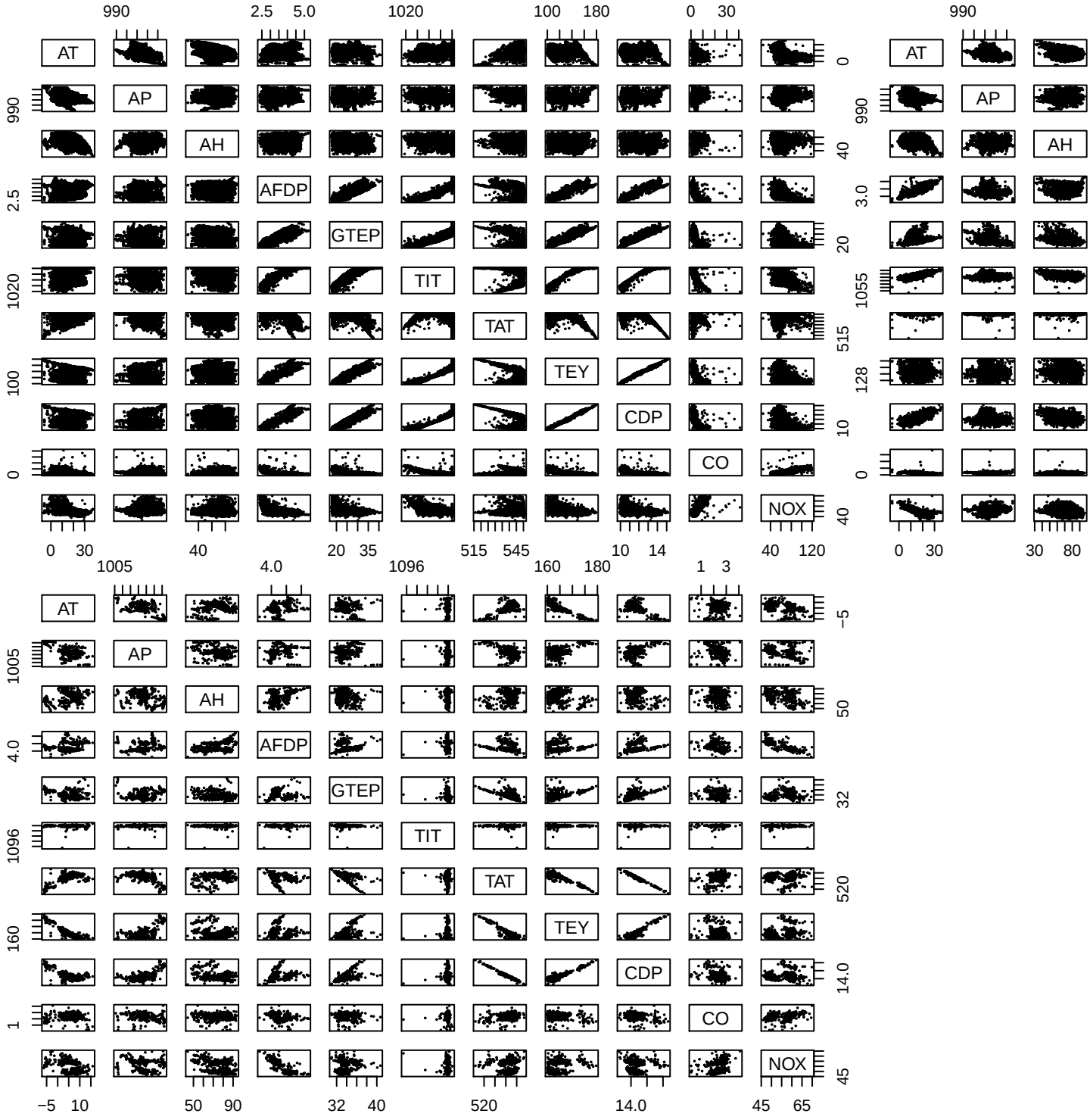


Figure 2:

Table 4: Pairwise Correlation Between Variables (All Data)

	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

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
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

Table 5: Pairwise Correlation Between Variables (Typical Energy Yield)

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
AT	1.00	-0.35	-0.33	0.84	0.28	0.79	0.01	-0.01	0.78	-0.45	-0.76
AP	-0.35	1.00	0.00	-0.19	-0.31	-0.15	-0.06	-0.03	-0.14	0.16	0.21
AH	-0.33	0.00	1.00	-0.12	-0.41	-0.34	0.01	-0.11	-0.33	-0.02	-0.14
AFDP	0.84	-0.19	-0.12	1.00	0.12	0.77	-0.02	0.09	0.75	-0.41	-0.69
GTEP	0.28	-0.31	-0.41	0.12	1.00	0.32	-0.04	0.29	0.32	0.00	0.03
TIT	0.79	-0.15	-0.34	0.77	0.32	1.00	0.19	0.52	0.91	-0.40	-0.56
TAT	0.01	-0.06	0.01	-0.02	-0.04	0.19	1.00	-0.02	-0.10	-0.08	-0.08
TEY	-0.01	-0.03	-0.11	0.09	0.29	0.52	-0.02	1.00	0.49	-0.03	0.11
CDP	0.78	-0.14	-0.33	0.75	0.32	0.91	-0.10	0.49	1.00	-0.37	-0.53
CO	-0.45	0.16	-0.02	-0.41	0.00	-0.40	-0.08	-0.03	-0.37	1.00	0.55
NOX	-0.76	0.21	-0.14	-0.69	0.03	-0.56	-0.08	0.11	-0.53	0.55	1.00

Table 6: Pairwise Correlation Between Variables (High Energy Yield)

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
AT	1.00	-0.52	0.20	0.24	-0.29	0.03	0.67	-0.89	-0.70	0.01	-0.42
AP	-0.52	1.00	-0.22	0.23	0.11	0.07	-0.55	0.55	0.59	-0.18	-0.22
AH	0.20	-0.22	1.00	0.30	-0.21	-0.01	0.23	-0.24	-0.24	-0.03	-0.34
AFDP	0.24	0.23	0.30	1.00	-0.08	0.00	-0.17	-0.05	0.15	-0.40	-0.82
GTEP	-0.29	0.11	-0.21	-0.08	1.00	-0.02	-0.59	0.51	0.56	-0.03	0.26
TIT	0.03	0.07	-0.01	0.00	-0.02	1.00	-0.04	0.06	0.09	0.06	-0.04
TAT	0.67	-0.55	0.23	-0.17	-0.59	-0.04	1.00	-0.92	-0.99	0.29	-0.06
TEY	-0.89	0.55	-0.24	-0.05	0.51	0.06	-0.92	1.00	0.94	-0.15	0.27
CDP	-0.70	0.59	-0.24	0.15	0.56	0.09	-0.99	0.94	1.00	-0.26	0.06
CO	0.01	-0.18	-0.03	-0.40	-0.03	0.06	0.29	-0.15	-0.26	1.00	0.39
NOX	-0.42	-0.22	-0.34	-0.82	0.26	-0.04	-0.06	0.27	0.06	0.39	1.00

Remove variables that are highly correlated.

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

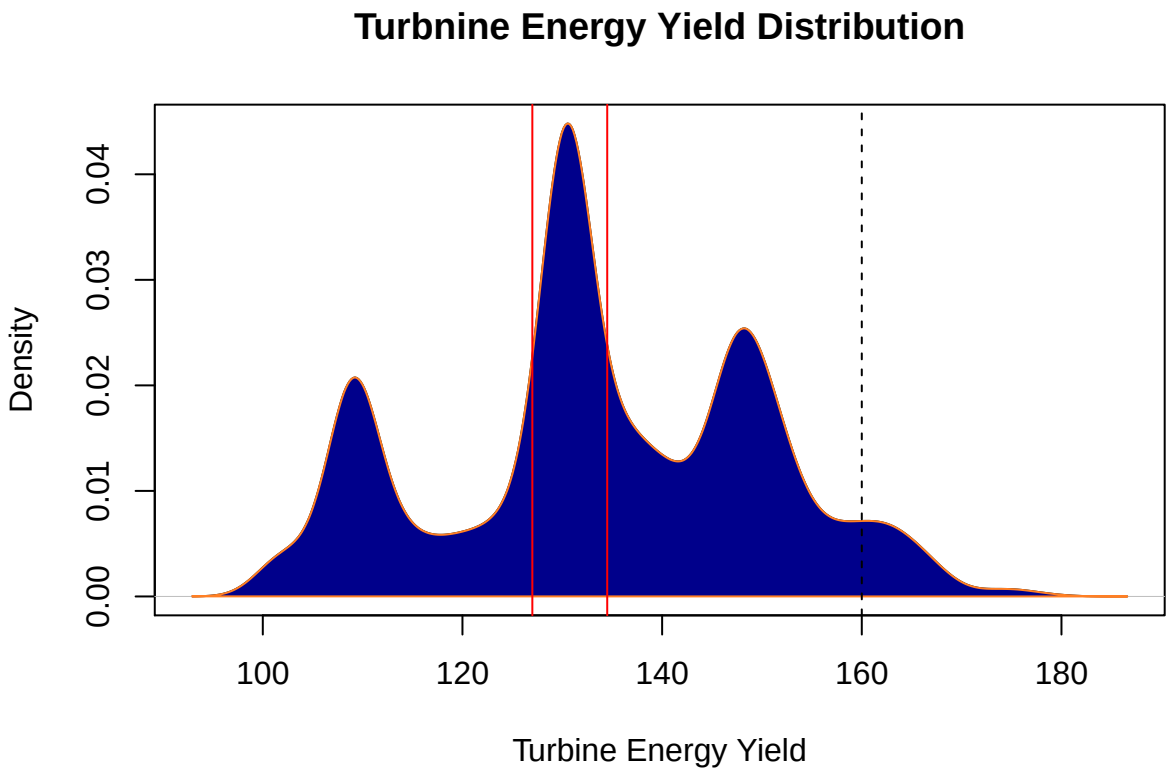
```
##       AP       AH     AFDP     GTEP      TAT      TEY      CDP
## 1.179915 1.406863 3.351823 1.496232 1.038742 1.836942 4.823918
```

```
##       AT       AP       AH     AFDP     GTEP      TIT      TAT
## 3.084771 1.948954 1.321587 1.996187 1.897431 1.027967 4.191605
```

Exploratory analysis shows possible linear relationships between the response variable CO and the feature variables CDP, TEY, TIT, GTEP and AFDP. Collinearity between some of the feature variables (TIT, CDP, and TEY) could cause some problems in our analysis and will likely lead to the removal of the redundant variables.

#density

```
d <- density(gt_2015$TEY)
plot(d, xlab = "Turbine Energy Yield", ylab = "Density", main = "Turbnine Energy Yield Distribution")
polygon(d, col = "blue4", border = "chocolate1")
abline(v = c(127,134.5,160), lty = c(1,1,2), col = c("red","red","black"))
```



Find and fix/remove ouliers

```
# plot(gt_2015$CO)
# abline(h = 15)

# which(gt_2015$CO > 15)
#
# gt_2015[c(seq(119-6,119+4)),]

gt_2015[c(1363,1364,1585,3977,3978,4762,5752,5753,6901),10] <- gt_2015[c(1363,1364,1585,3977,3978,4762,5

# gt_2015[c(1363,1364,1585,3977,3978,4762,5752,5753,6901),]


gt_2015 <- gt_2015[-c(1796,1713,1712,1711,1710,1709,1301,1009,626,121,120,119,118,117,116,115),]
```

```r
# which(gt_2015$CO > 15)

# plot(gt_2015$CO)

gt_2015_typical <- gt_2015[gt_2015$TEY <= 134.5 & gt_2015$TEY >= 127,]
gt_2015_high <- gt_2015[gt_2015$TEY >= 160,]
```

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 β 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.

Linear and Lasso stepwise AIC Models

#All Data

#Typical Energy Yield (127-134.5)

#High Energy Yield (160+)

#Box-Cox Lambdas

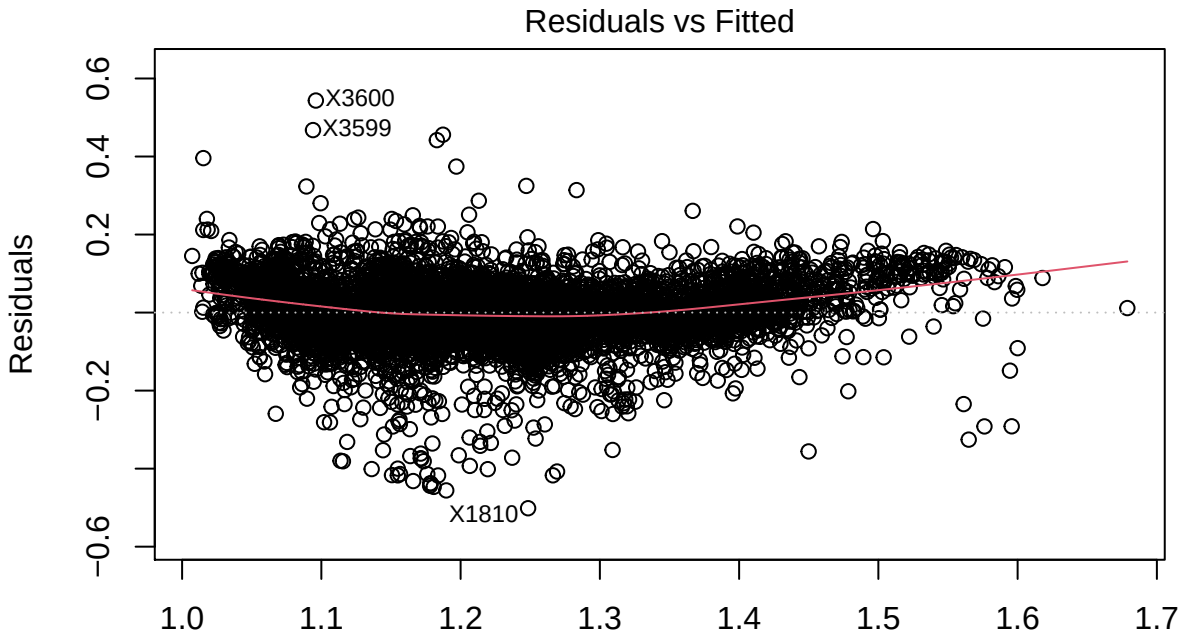
```r
all_lambda <- boxcox(all_linear_mod_lm, plotit = FALSE)$x[which.max(boxcox(all_linear_mod_lm, plotit = 
typical_lambda <- boxcox(typical_linear_mod_lm, plotit = FALSE)$x[which.max(boxcox(typical_linear_mod_l
high_lambda <- boxcox(high_linear_mod_lm, plotit = FALSE)$x[which.max(boxcox(high_linear_mod_lm, plotit
```

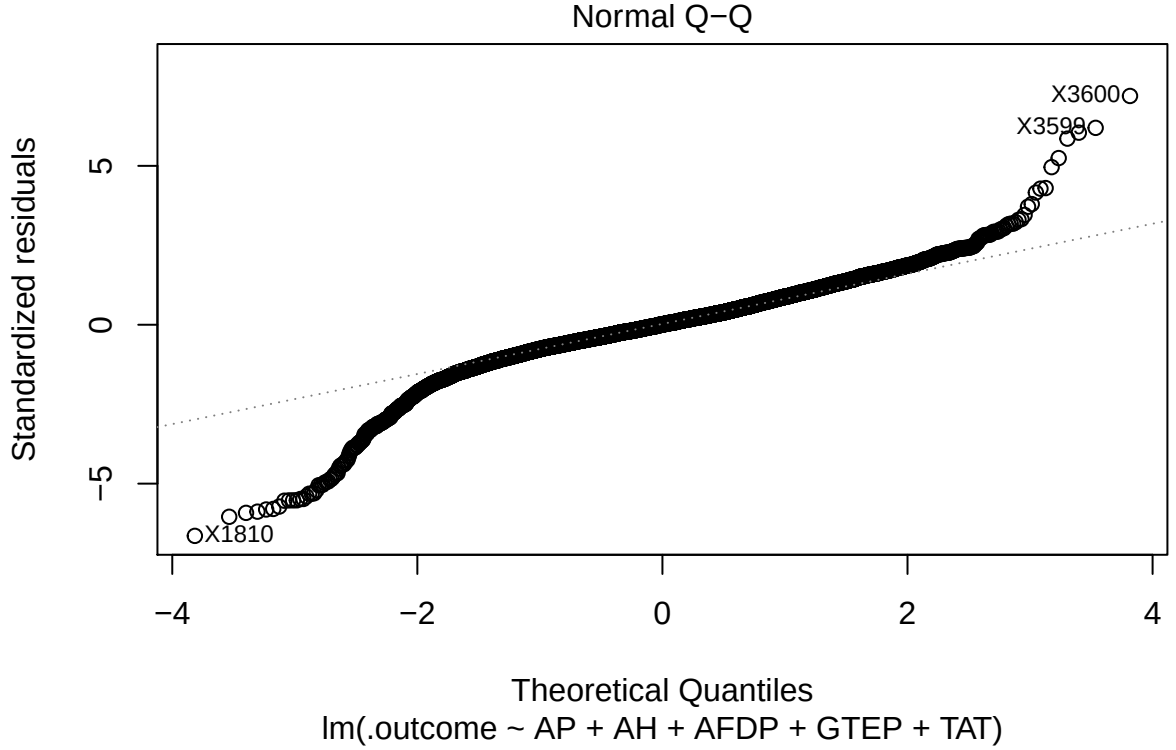
#Box-Cox Transformed Models

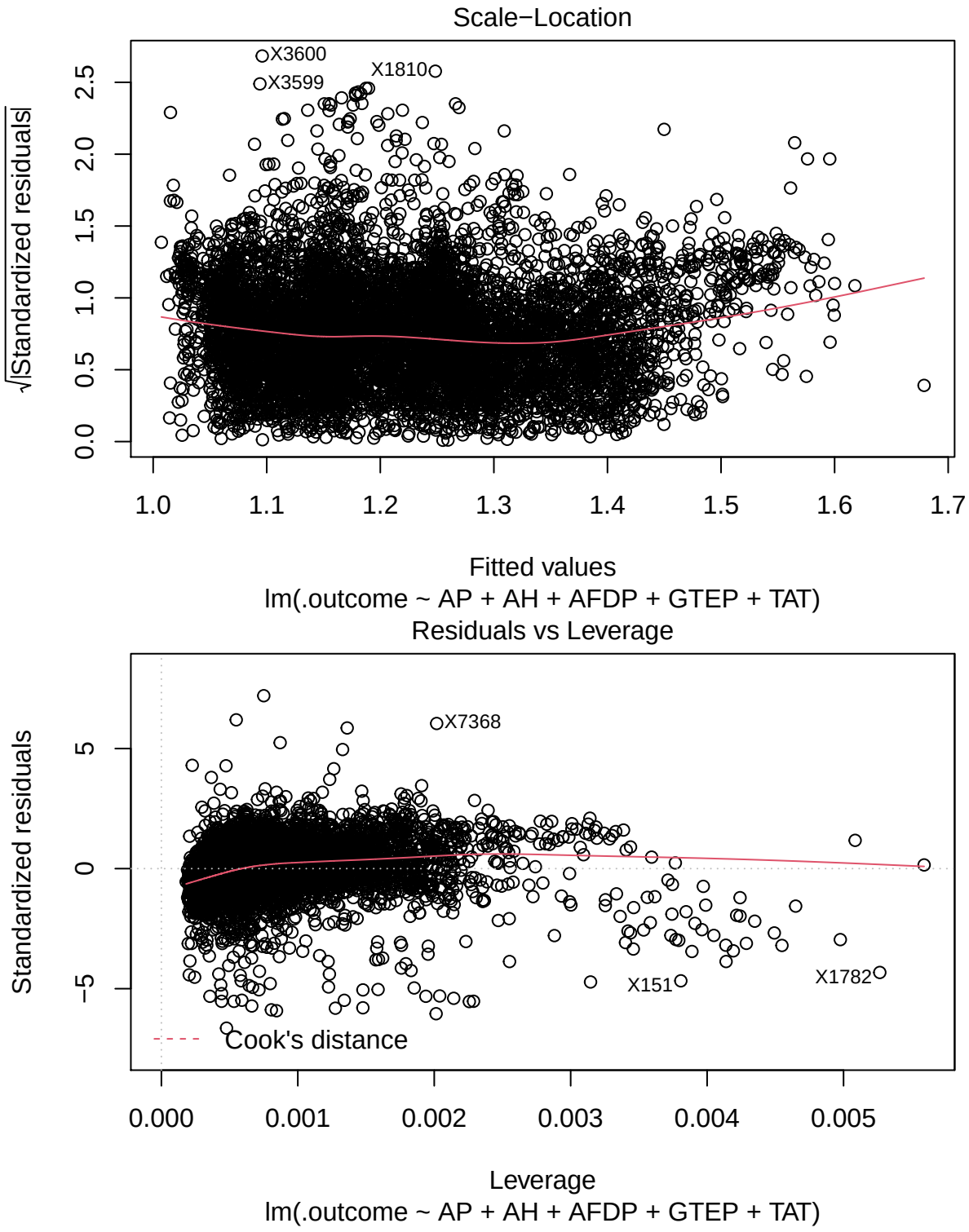
```r
#All Data Box-Cox Transformed Linear Model
set.seed(10)
all_bc_linear_mod <- train(
          form = CO^(.2) ~ . - NOX  - TIT - CDP - TEY ,
          data = gt_2015,
          method = "lmStepAIC",
          trControl = cv_5,
          nvmax = 10,
          trace = FALSE
)
```

[1] Dalpiaz David, Applied Statistics in R, https://urldefense.com/v3/__https://daviddalpiaz.github.io/appliedstats/model-diagnostics.html__;!!DZ3fjg!pZQU6uJCClJrohh1D9pa0MjKMi32lYqIPLCJl_vX4wq1QtzWTjyE-7kqGzwyrgx2FQEo$

```r
all_bc_linear_mod_lm <- lm(CO^(.2) ~ AT + AP + AH + AFDP + GTEP + TAT, data = gt_2015)
plot(all_bc_linear_mod$finalModel)
```

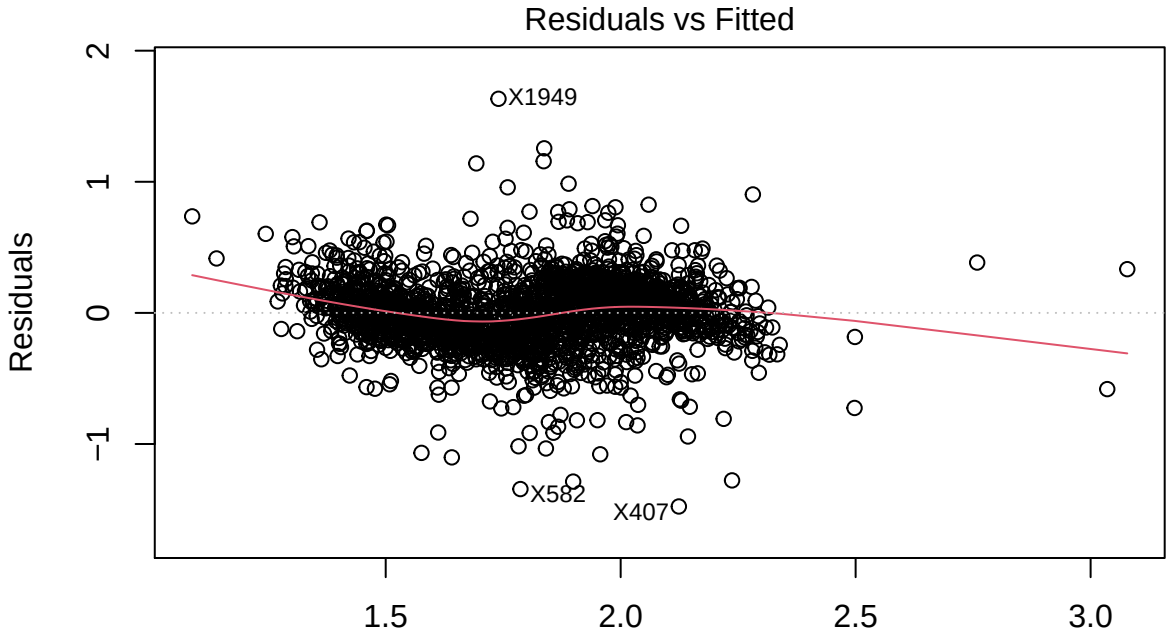


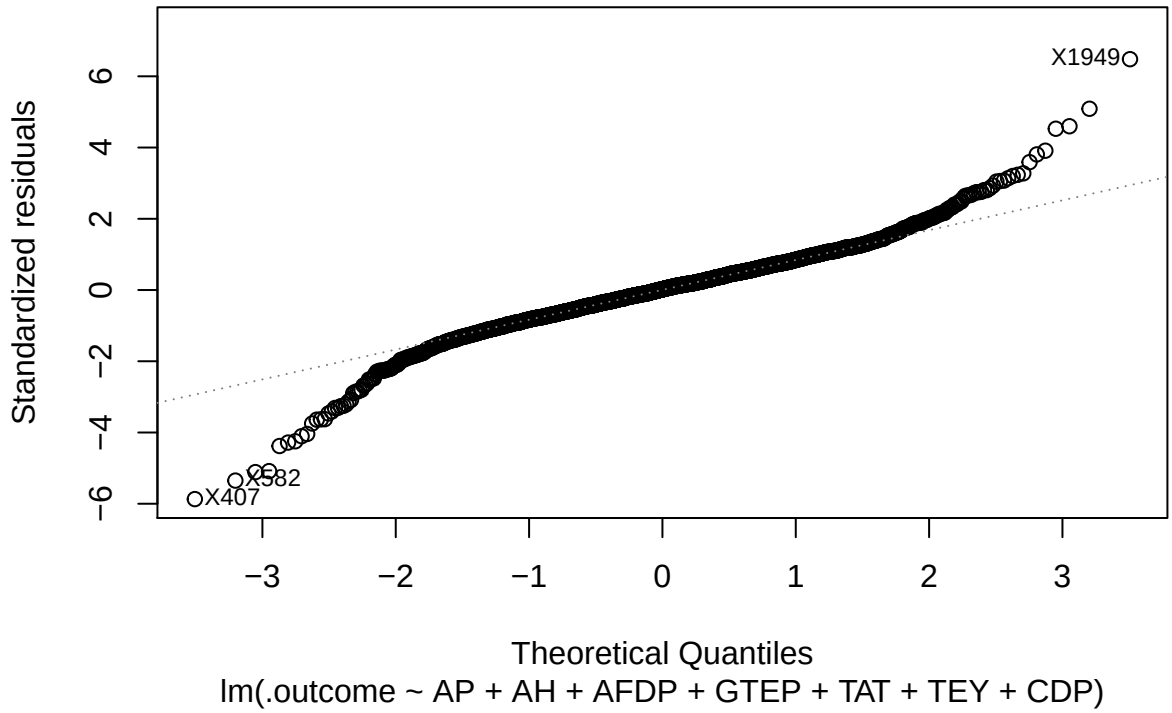


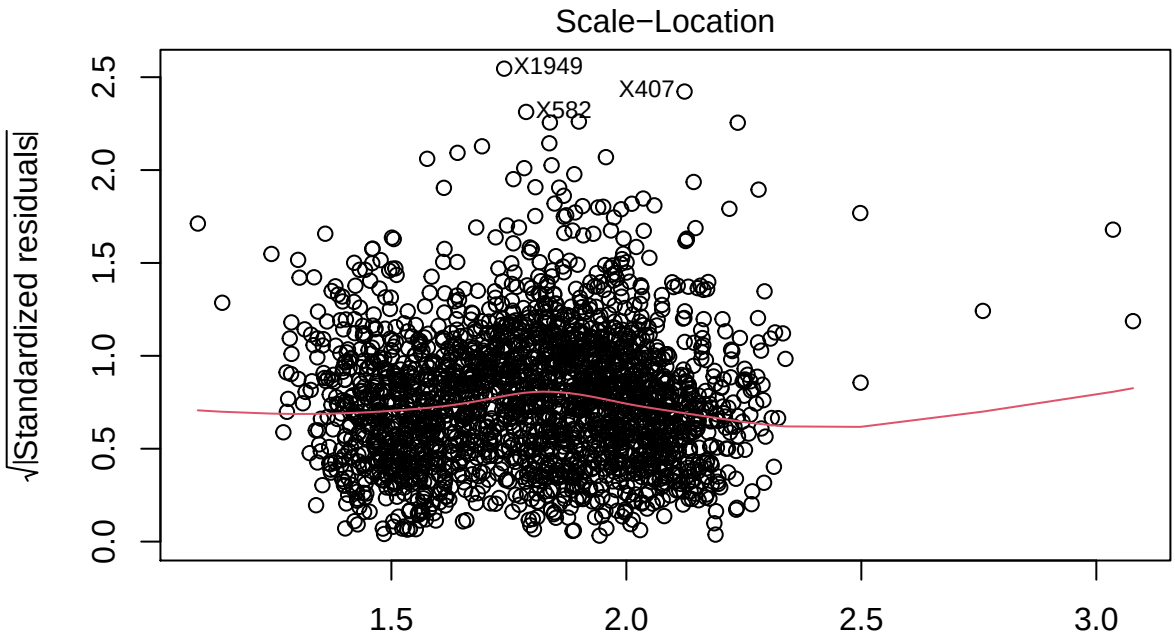


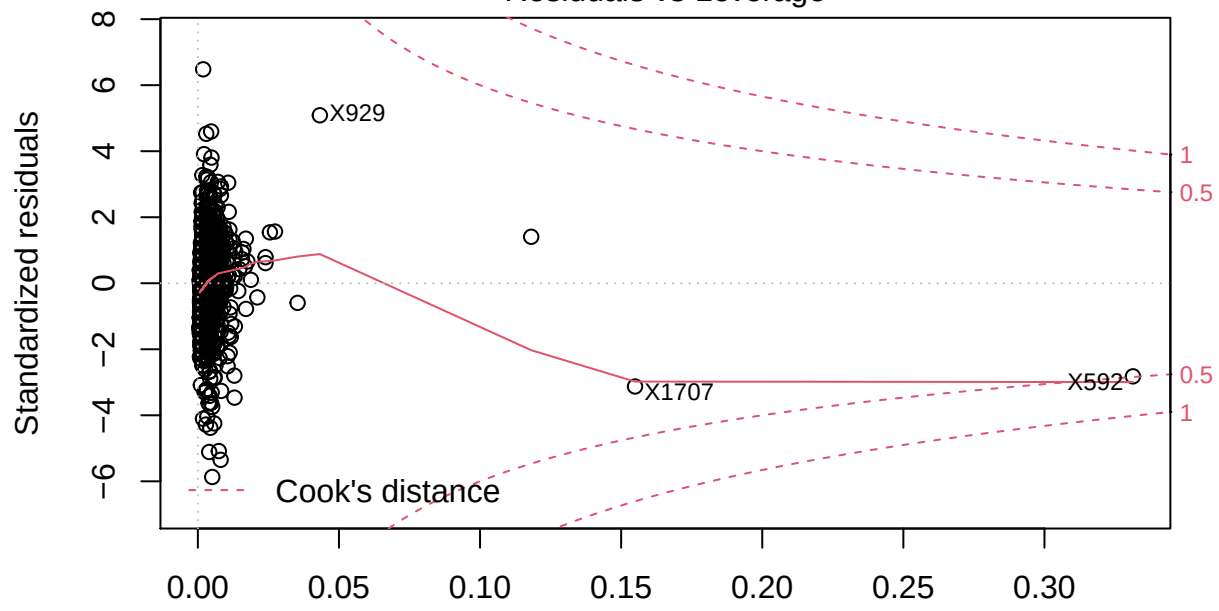

```
#Typical Data Box-Cox Transformed Linear Model
typical_bc_linear_mod <- train(
          form = CO^(.6) ~ . - NOX - TIT - AT,
          data = gt_2015_typical,
          method = "lmStepAIC",
          trControl = cv_5,
```

```
          nvmax = 10,
          trace = FALSE
)
typical_bc_linear_mod_lm <- lm(CO^(.6) ~ AP + AH + AFDP + GTEP + TAT + TEY + CDP, data = gt_2015_typical
plot(typical_bc_linear_mod$finalModel)
```



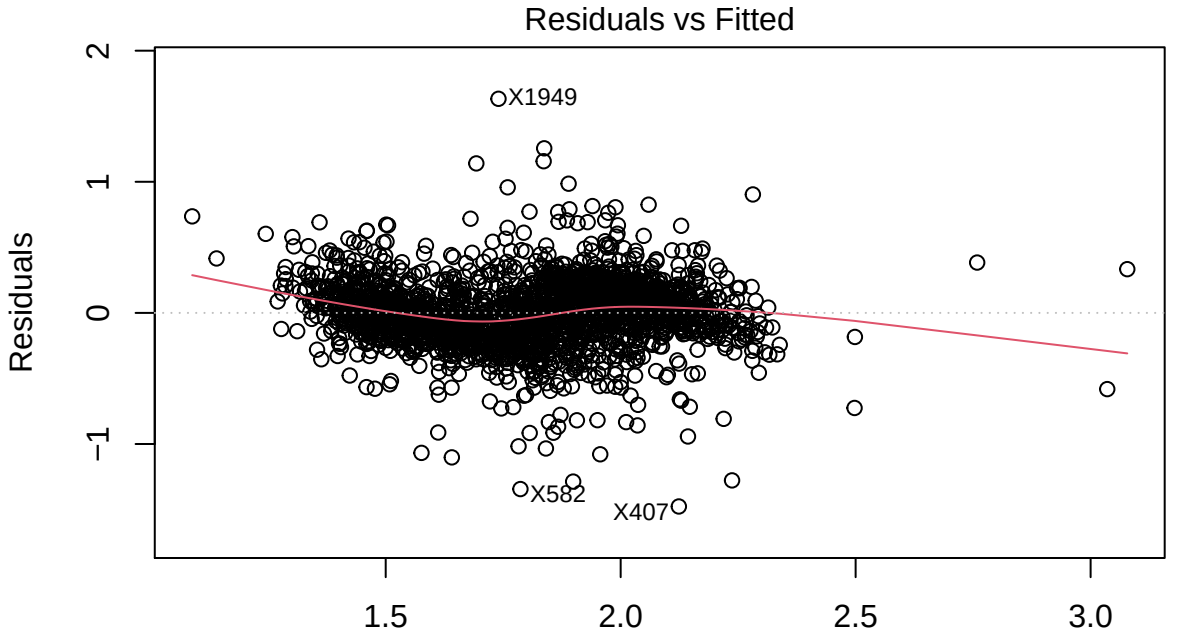


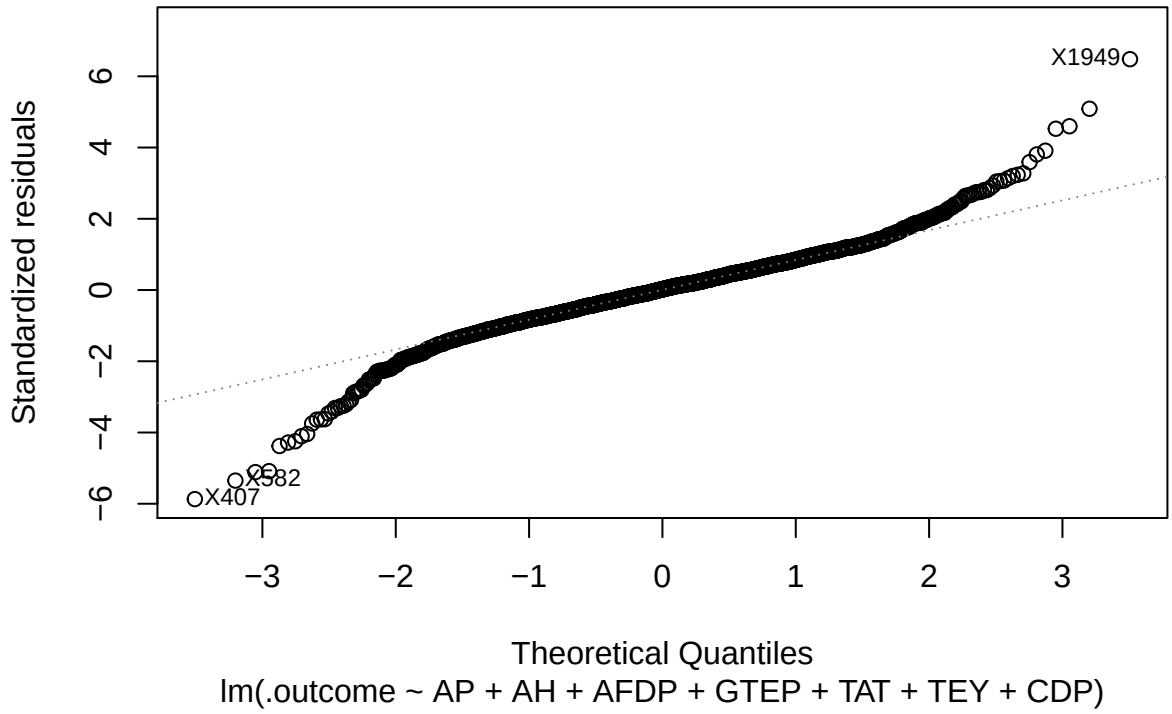


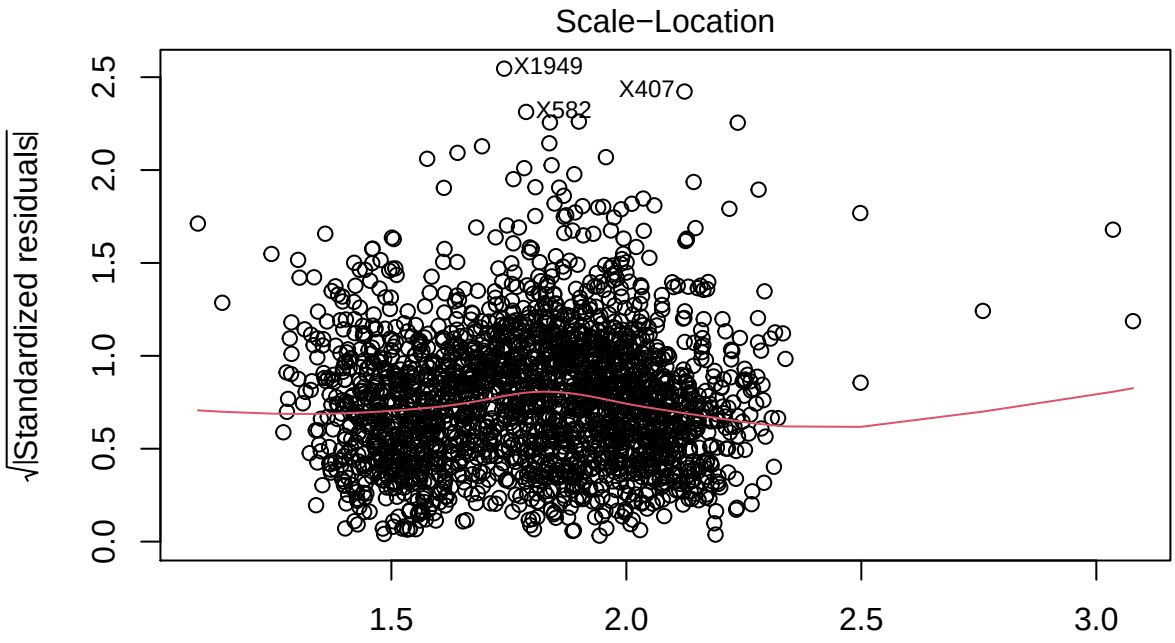


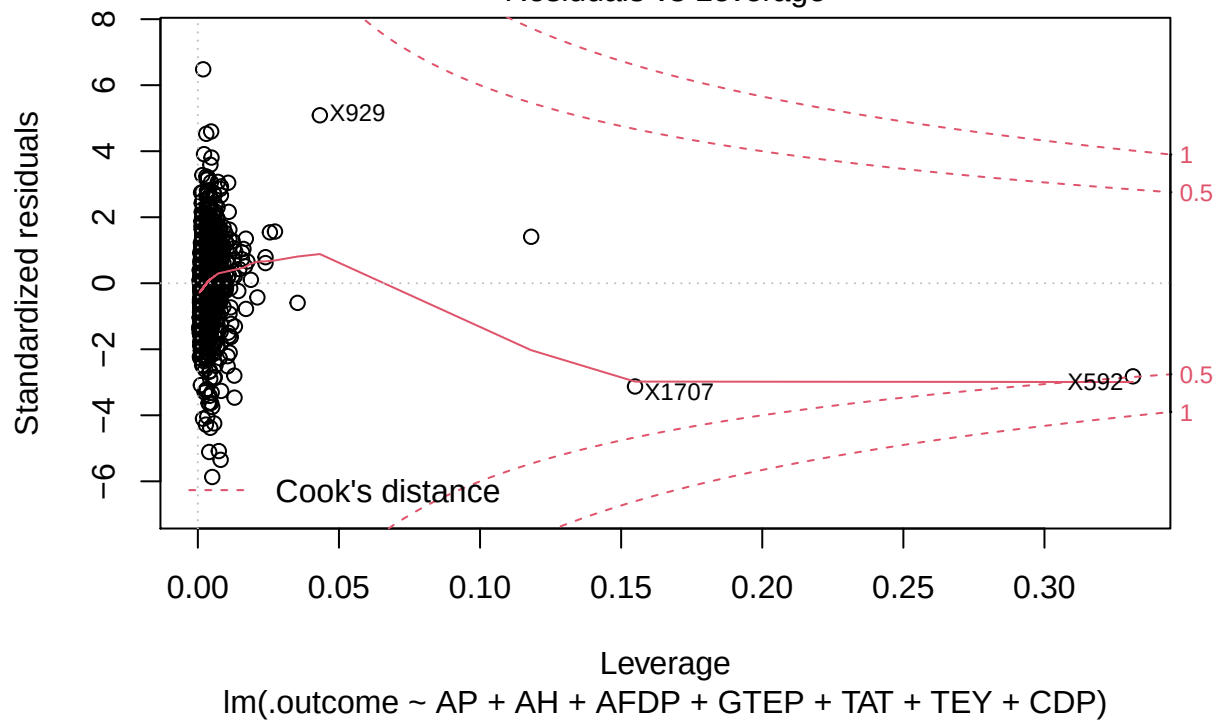

```
#High Data Box-Cox Transformed Linear Model
high_bc_linear_mod <- train(
          form = CO^(1.9) ~ . - NOX - TEY - CDP,
          data = gt_2015_high,
          method = "lmStepAIC",
          trControl = cv_5,
```

```
          nvmax = 10,
          trace = FALSE
)
high_bc_linear_mod_lm <- lm(CO^(1.9) ~ AT + AFDP + GTEP + TIT + TAT, data =  gt_2015_high)
plot(typical_bc_linear_mod$finalModel)
```









#Results

```
all_linear_mod$results
```

```
##   parameter     RMSE  Rsquared       MAE     RMSESD RsquaredSD    MAESD
## 1      none 1.108804 0.6866739 0.7753151 0.04693766 0.01430024 0.015165
```

```
all_bc_linear_mod$results
```

```
##   parameter       RMSE  Rsquared        MAE      RMSESD  RsquaredSD       MAESD
## 1      none 0.07555509 0.6962445 0.05400183 0.002330428 0.007666233 0.001047801
```

```
all_lasso_mod$results
```

```
##   fraction     RMSE  Rsquared       MAE     RMSESD RsquaredSD       MAESD
## 1      0.1 1.727066 0.5006197 1.1283432 0.06268661 0.01278596 0.023831343
## 2      0.5 1.242282 0.6375583 0.7741334 0.05267314 0.01076525 0.020973024
## 3      0.9 1.111207 0.6851262 0.7667759 0.02793658 0.01079260 0.005485551
```

```
typical_linear_mod$results
```

```
##   parameter      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1      none 0.6292529 0.4781443 0.4559889 0.01866185 0.04420794 0.00576473
```

```
typical_bc_linear_mod$results
```

```
##   parameter      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1      none 0.2527683 0.4756394 0.1843201 0.01709637 0.03971585 0.00984201
```

```
typical_lasso_mod$results
```

```
##   fraction      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1      0.1 0.7938734 0.3848690 0.6493829 0.05224618 0.04220159 0.02225905
## 2      0.5 0.6593323 0.4354231 0.4903654 0.05993301 0.04645829 0.02962297
## 3      0.9 0.6292803 0.4780133 0.4587284 0.05468789 0.04486680 0.02756426
```

```
high_linear_mod$results
```

```
##   parameter     RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1      none 0.442826 0.2248008 0.3183321 0.04574879 0.05330011 0.02670168
```

```
high_bc_linear_mod$results
```

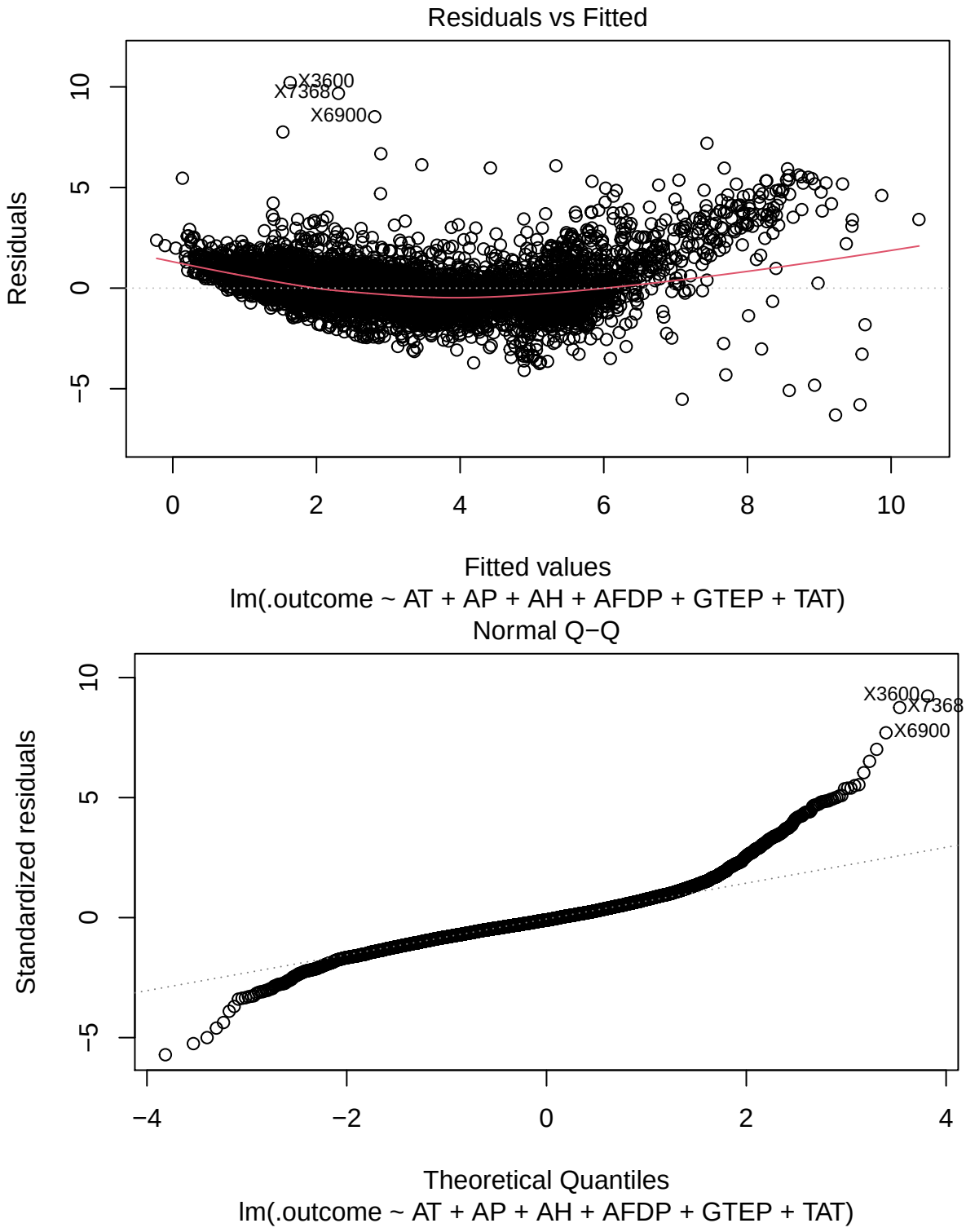
```
##   parameter     RMSE  Rsquared      MAE    RMSESD RsquaredSD      MAESD
## 1      none 1.688899 0.2089451 1.248995 0.1505623 0.09514931 0.08725389
```

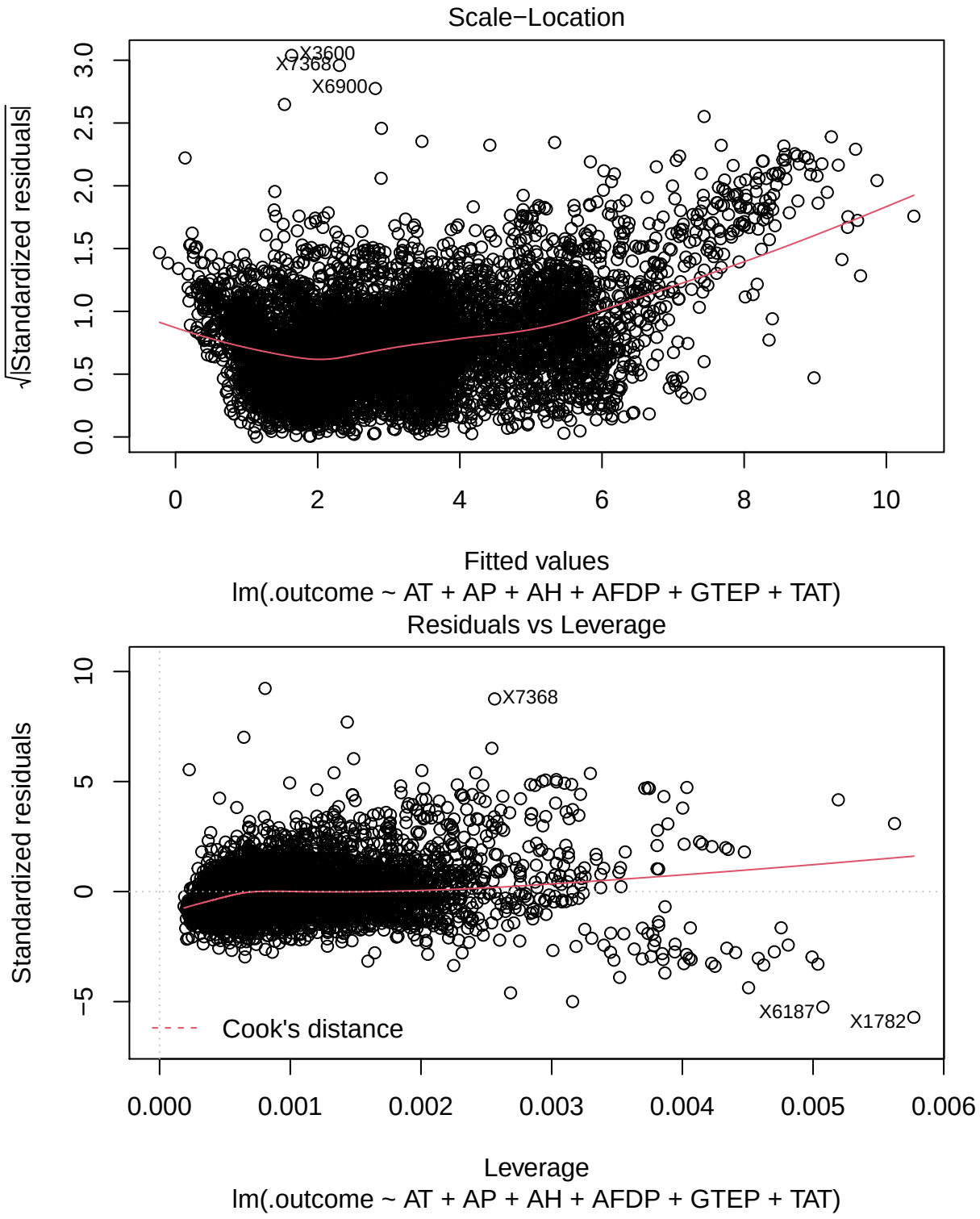
```
high_lasso_mod$results
```

```
##   fraction      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1      0.1 0.4772263 0.1780640 0.3478409 0.04551584 0.09297049 0.02313481
## 2      0.5 0.4483691 0.2082968 0.3171812 0.05022966 0.07271164 0.03219795
## 3      0.9 0.4479087 0.2113017 0.3214103 0.04943640 0.08500324 0.03582348
```

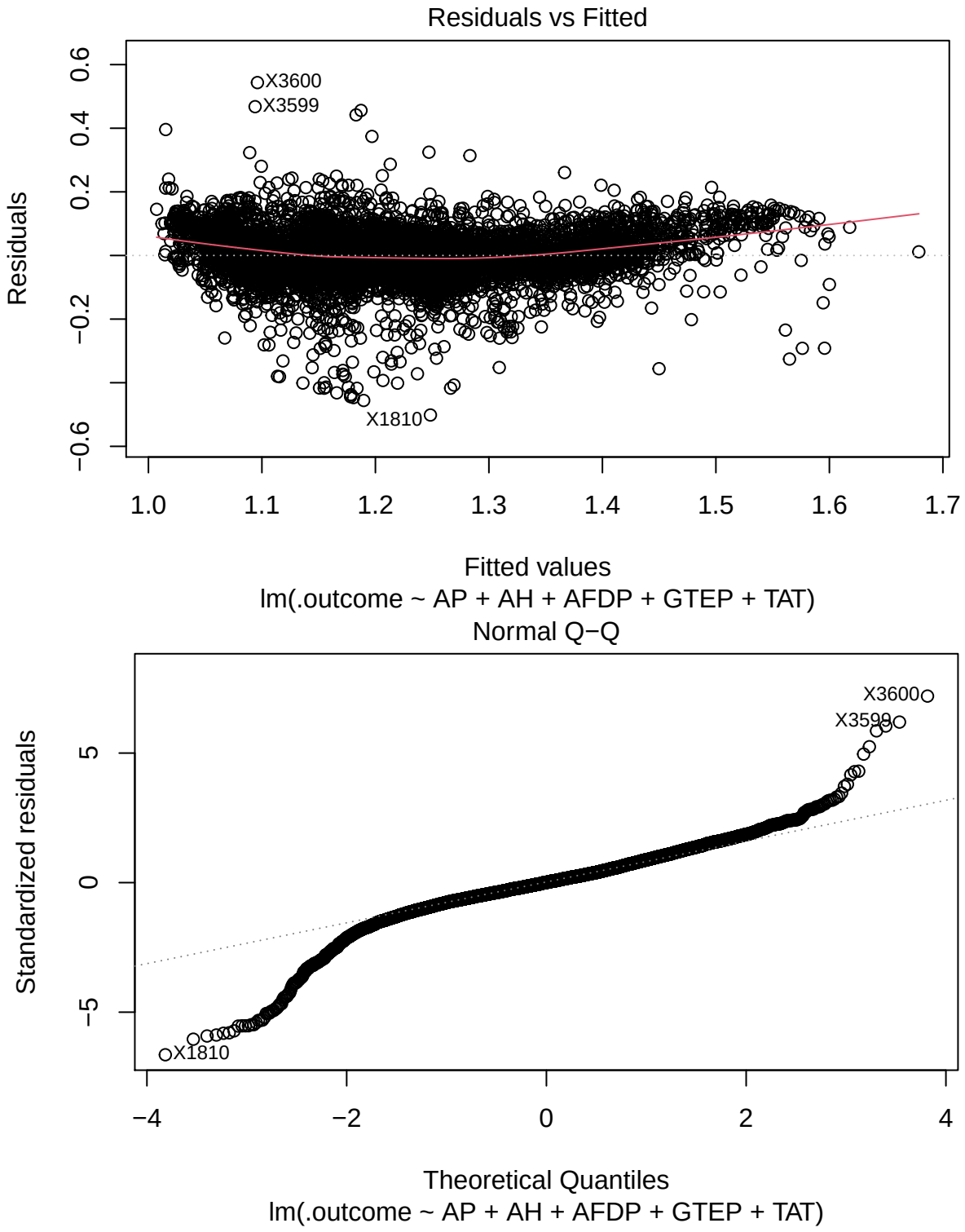
#plots

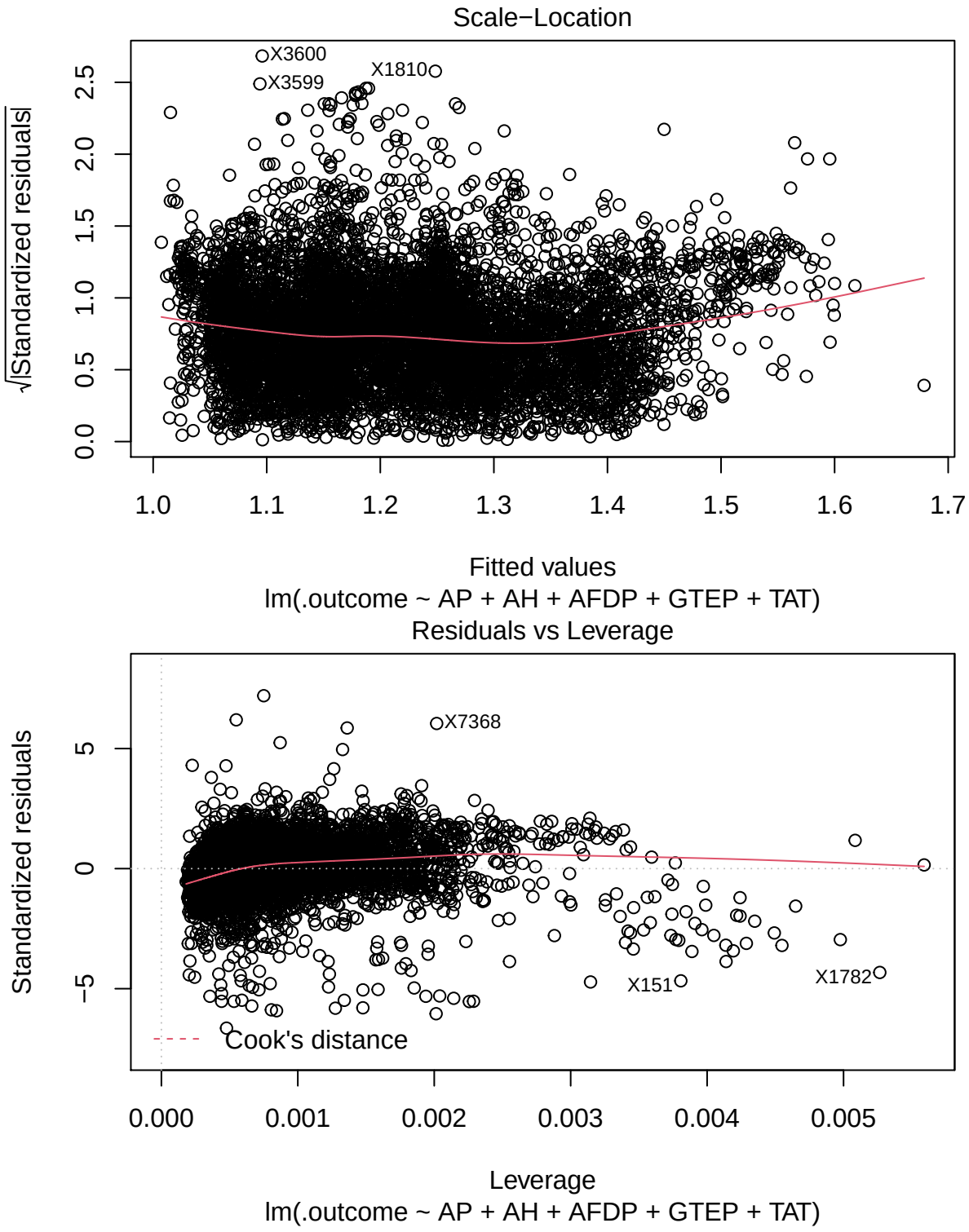
```
plot(all_linear_mod$finalModel)
```


Residuals vs Fitted
X3600
X7368
X6900
Residuals
10
5
0
−5
0
2
4
6
8
10
Fitted values
lm(.outcome ~ AT + AP + AH + AFDP + GTEP + TAT)
Normal Q−Q
X3600
X7368
X6900
Standardized residuals
10
5
0
−5
−4
−2
0
2
4
Theoretical Quantiles
lm(.outcome ~ AT + AP + AH + AFDP + GTEP + TAT)

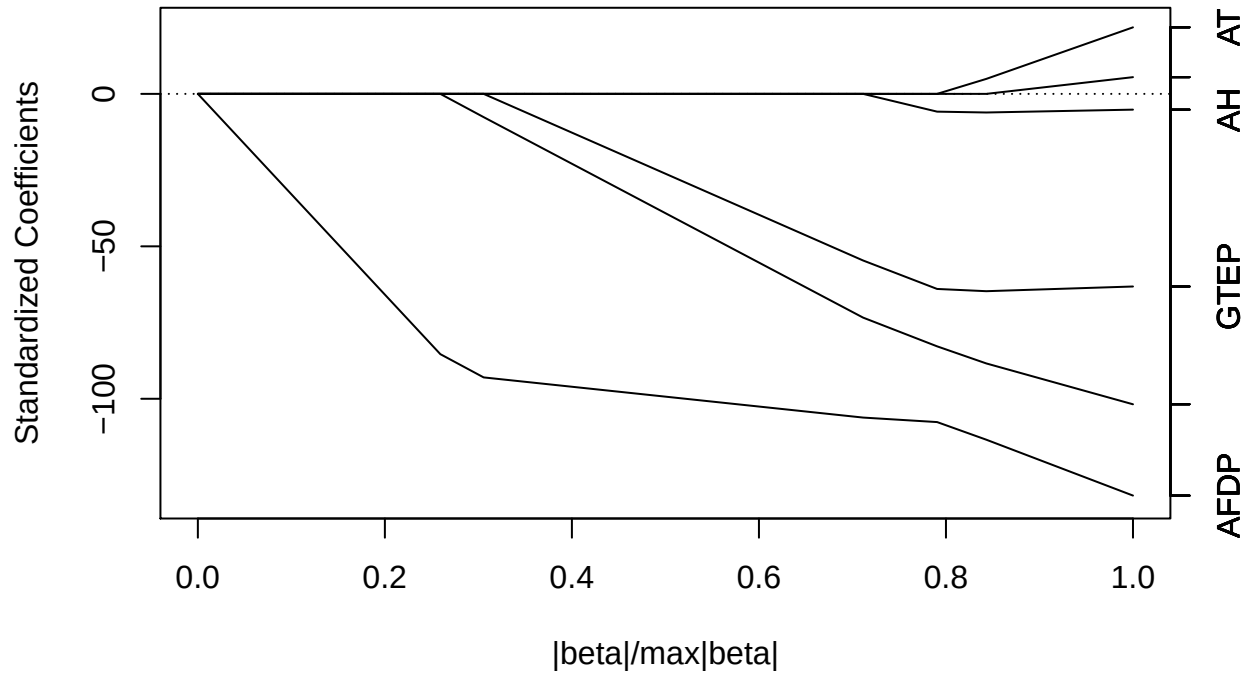



```
plot(all_bc_linear_mod$finalModel)
```

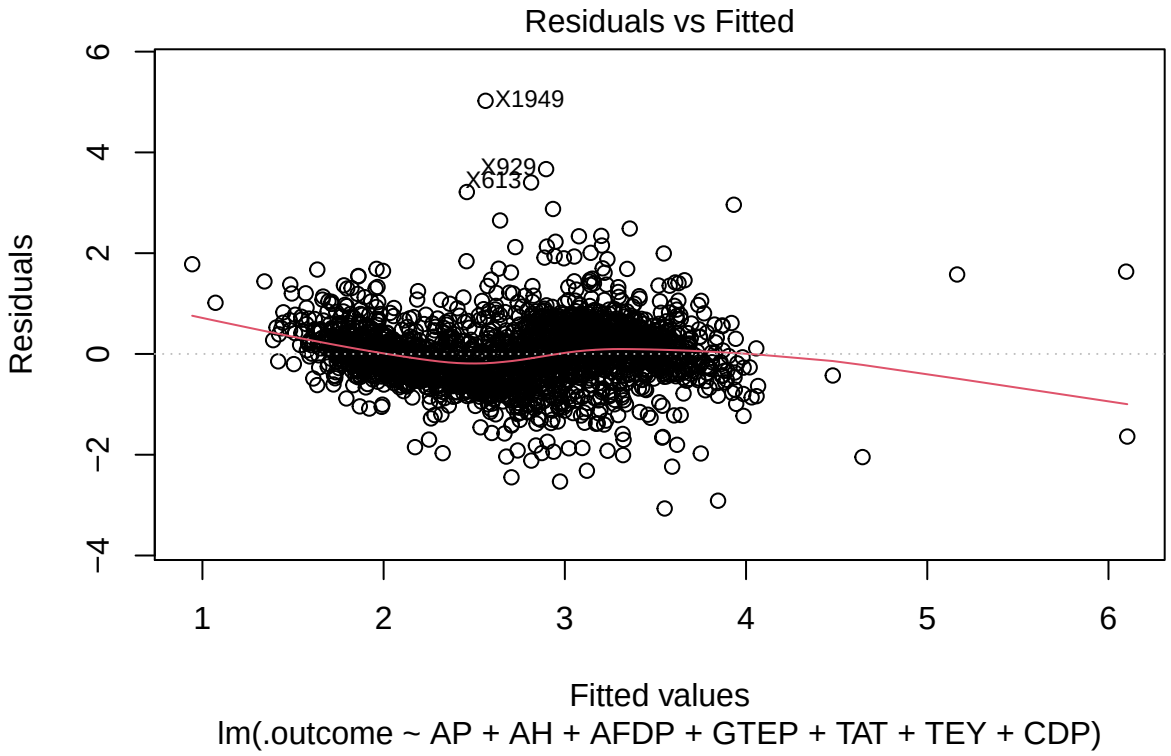

Residuals vs Fitted
Residuals
Fitted values
lm(.outcome ~ AP + AH + AFDP + GTEP + TAT)
X3600
X3599
X1810
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(.outcome ~ AP + AH + AFDP + GTEP + TAT)
X3600
X3599
X1810

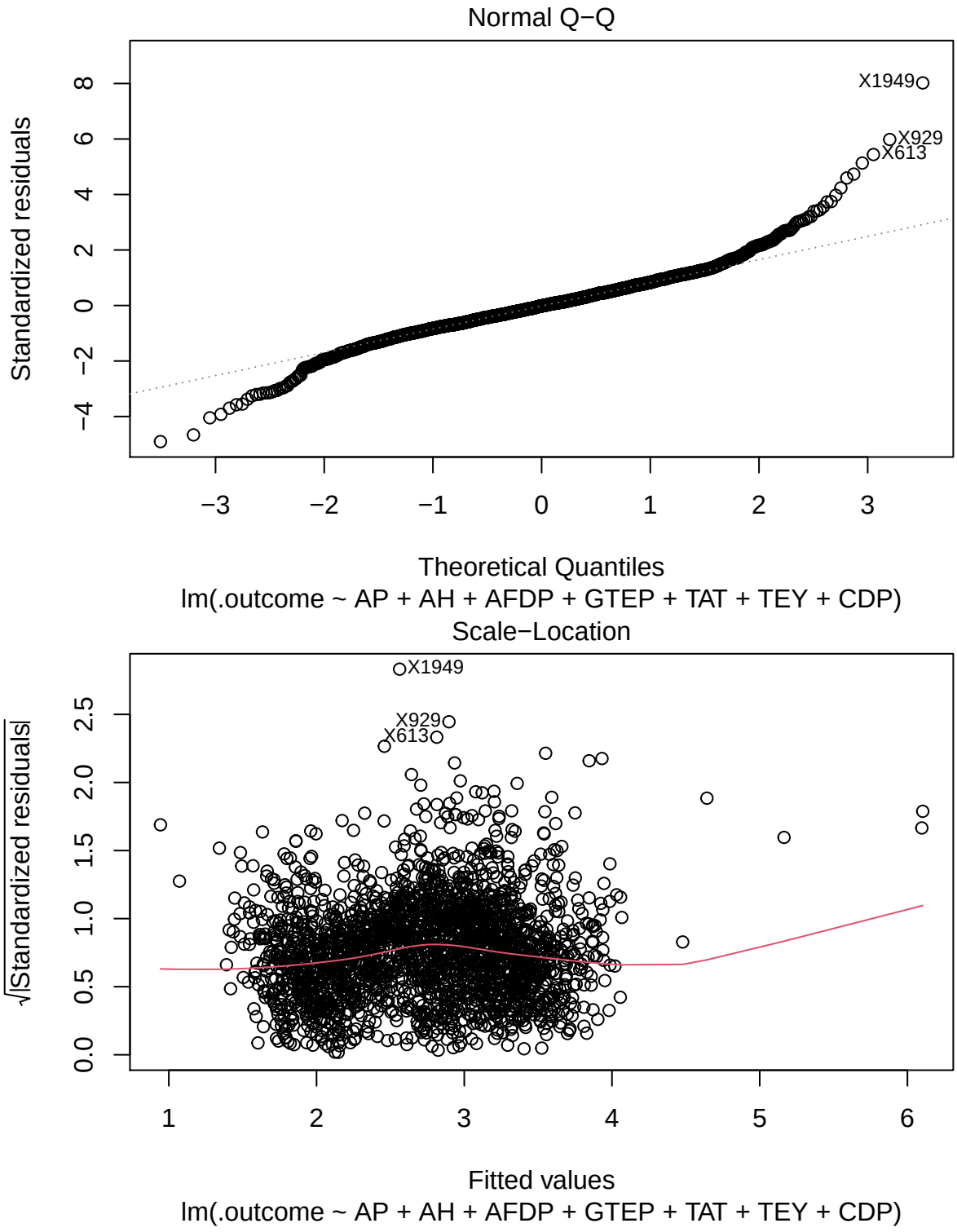


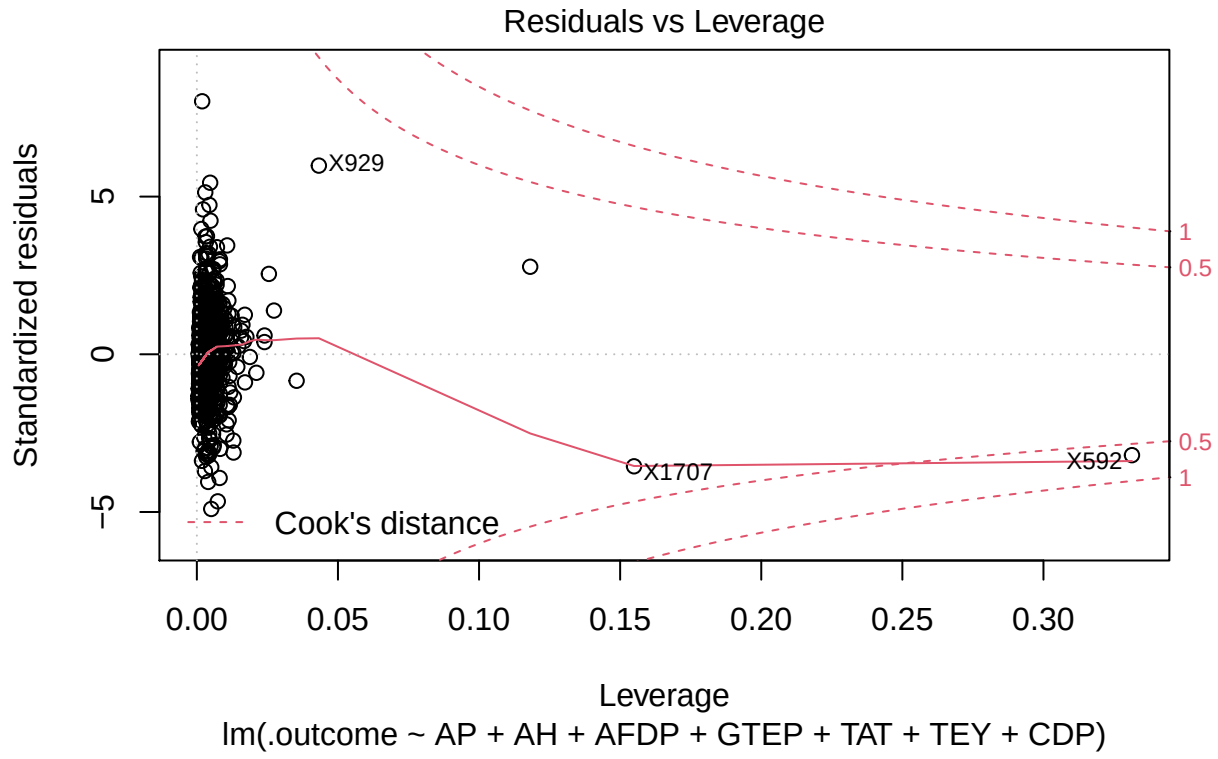

```
plot(all_lasso_mod$finalModel)
```



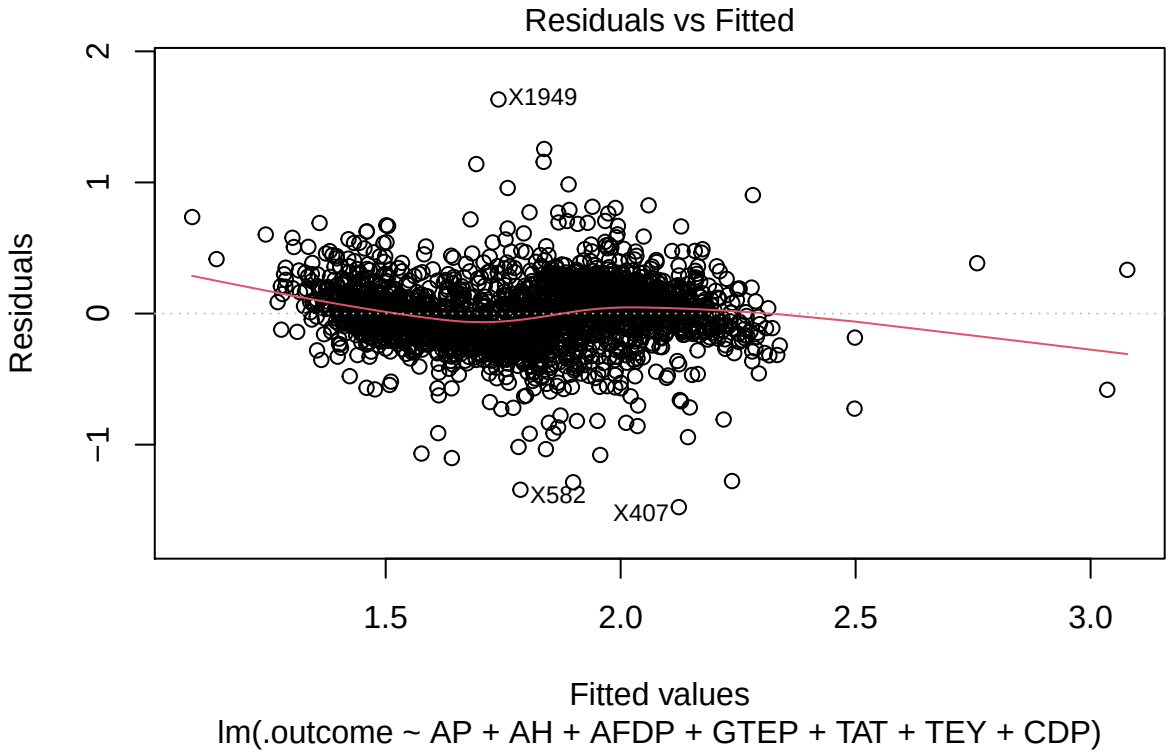

```
plot(typical_linear_mod$finalModel)
```

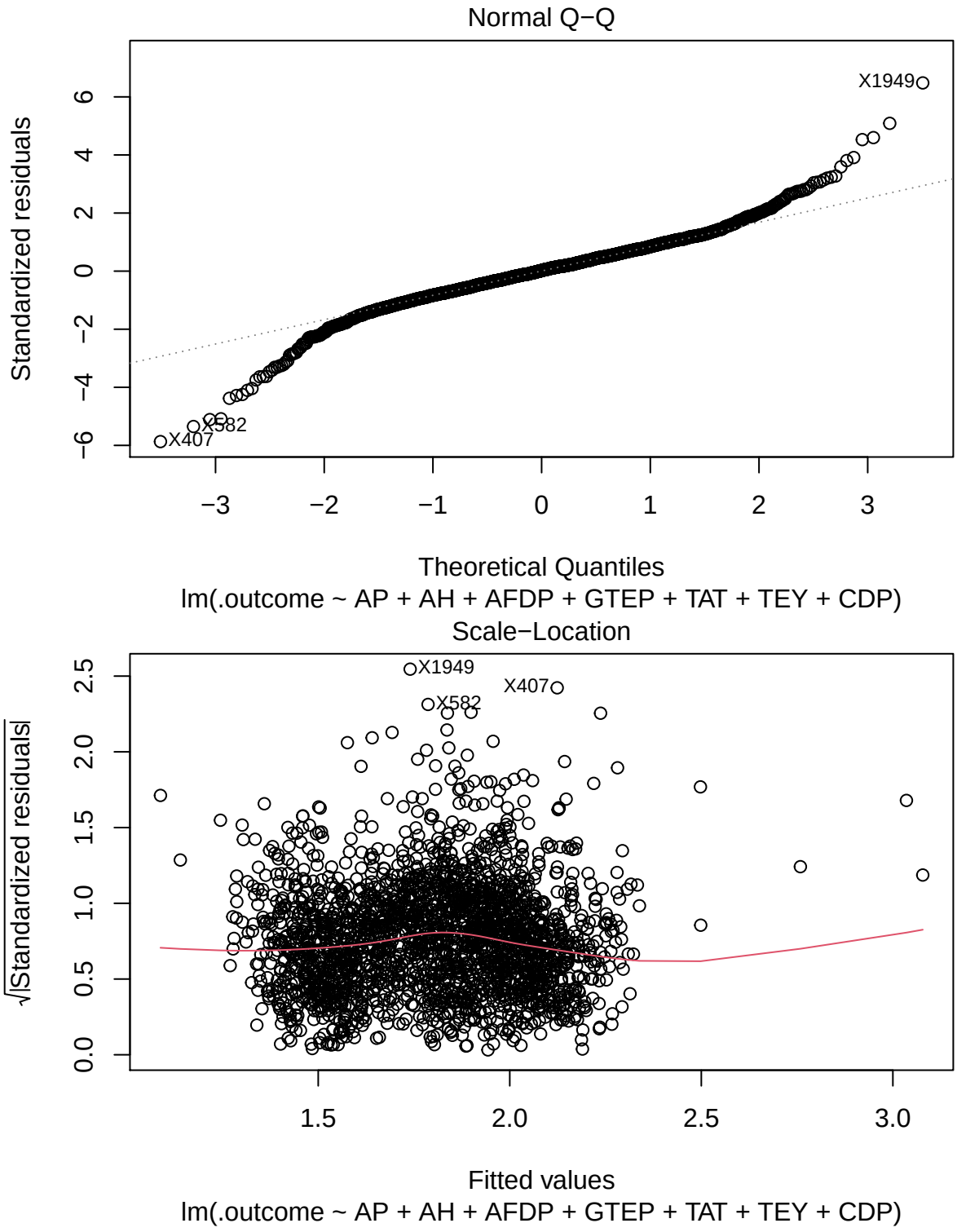


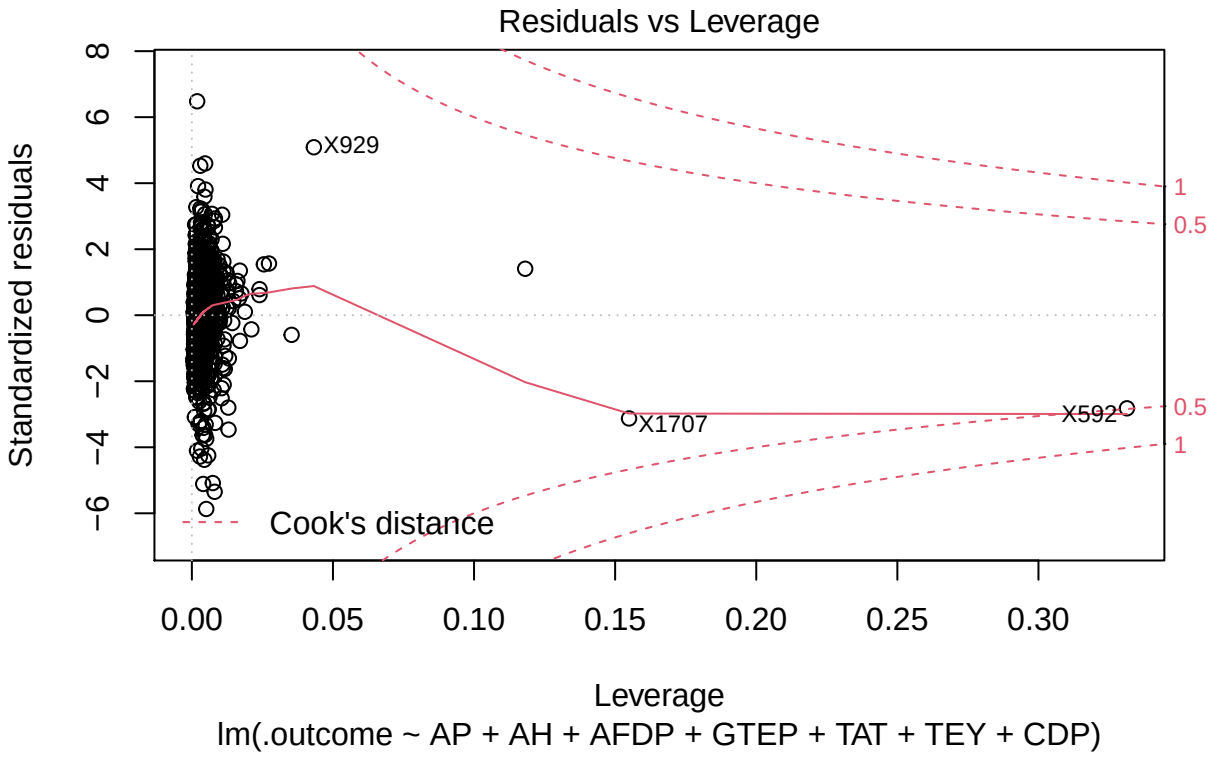

Normal Q−Q
Standardized residuals
X1949
X929
X613
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
6
8
Theoretical Quantiles
lm(.outcome ~ AP + AH + AFDP + GTEP + TAT + TEY + CDP)
Scale−Location
√|Standardized residuals|
X1949
X929
X613
1
2
3
4
5
6
0.0
0.5
1.0
1.5
2.0
2.5
Fitted values
lm(.outcome ~ AP + AH + AFDP + GTEP + TAT + TEY + CDP)



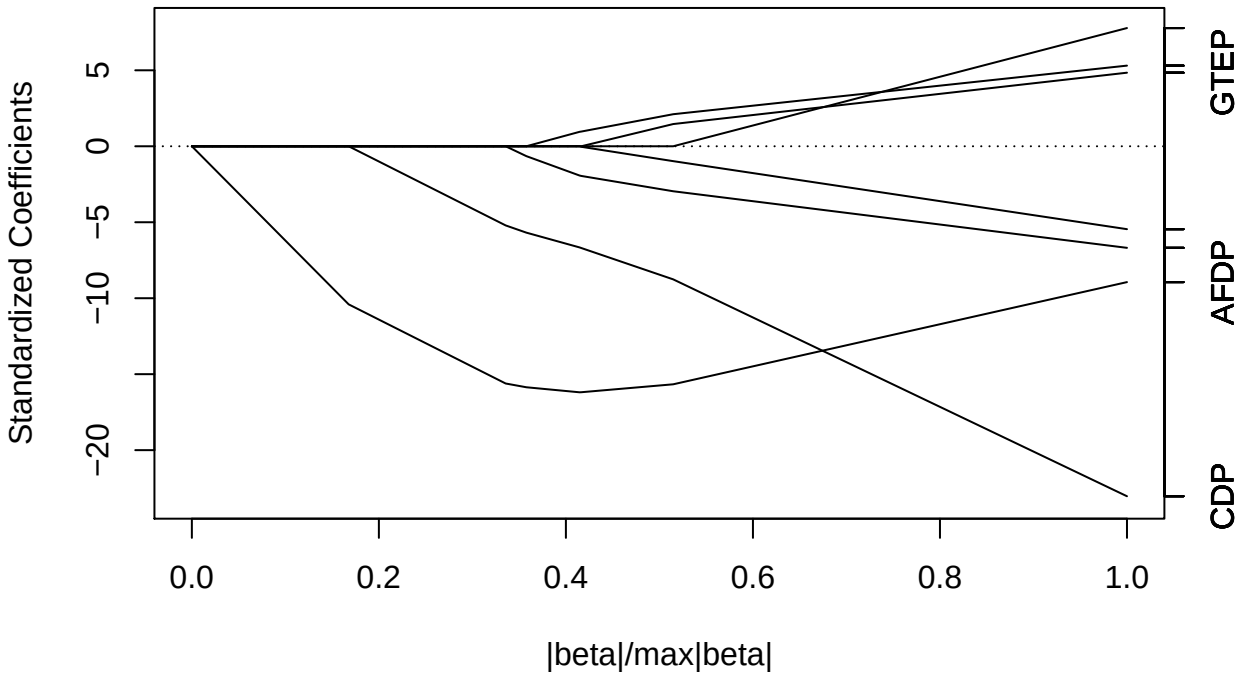

```
plot(typical_bc_linear_mod$finalModel)
```



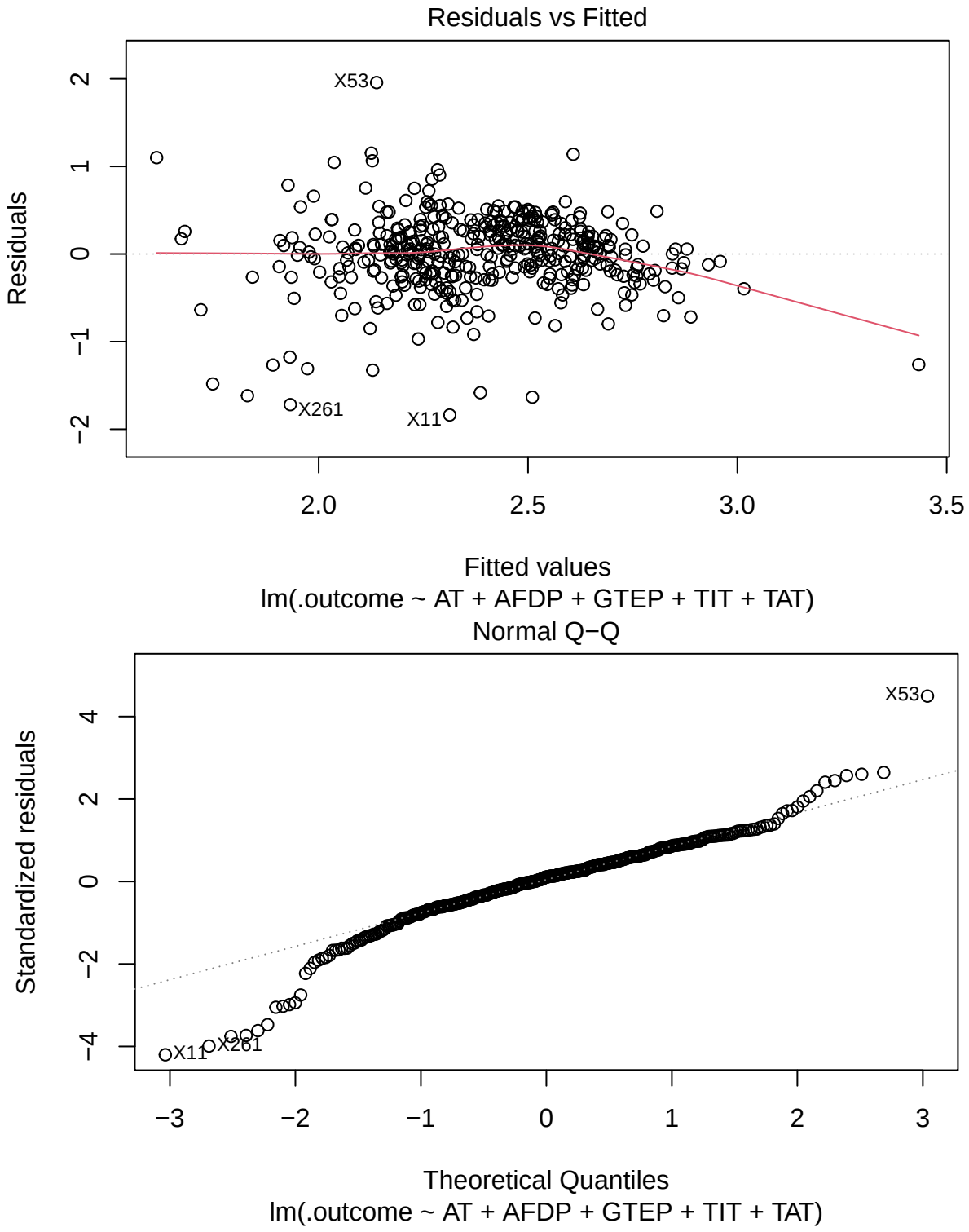

Normal Q−Q
Standardized residuals
X1949
X582
X407
Theoretical Quantiles
lm(.outcome ~ AP + AH + AFDP + GTEP + TAT + TEY + CDP)
Scale−Location
√|Standardized residuals|
X1949
X407
X582
Fitted values
lm(.outcome ~ AP + AH + AFDP + GTEP + TAT + TEY + CDP)

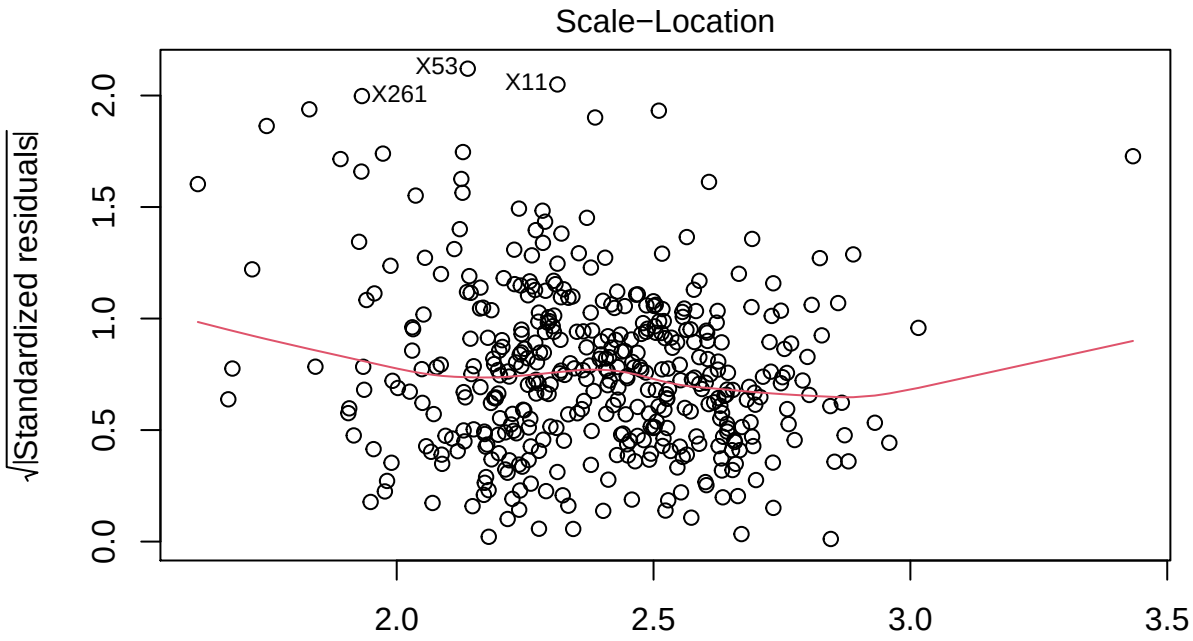


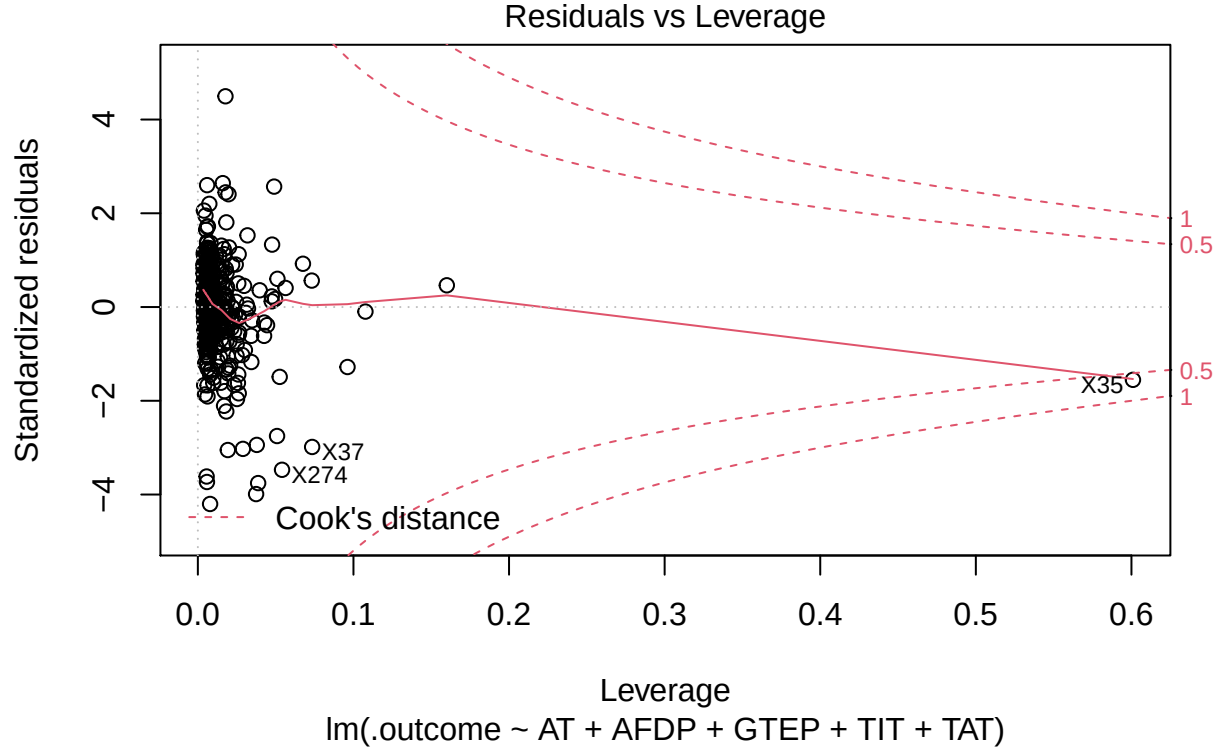

```
plot(typical_lasso_mod$finalModel)
```



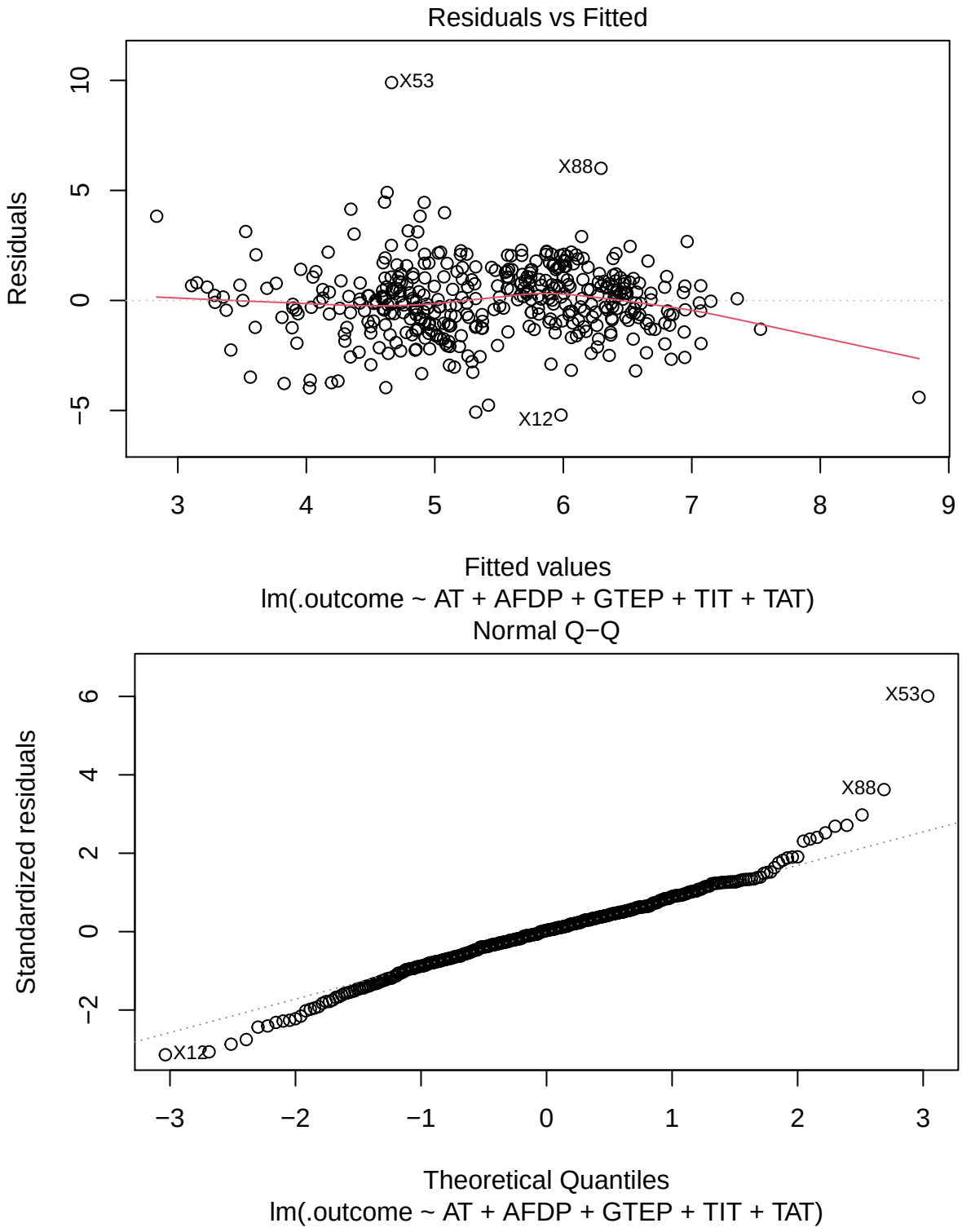

```
plot(high_linear_mod$finalModel)
```

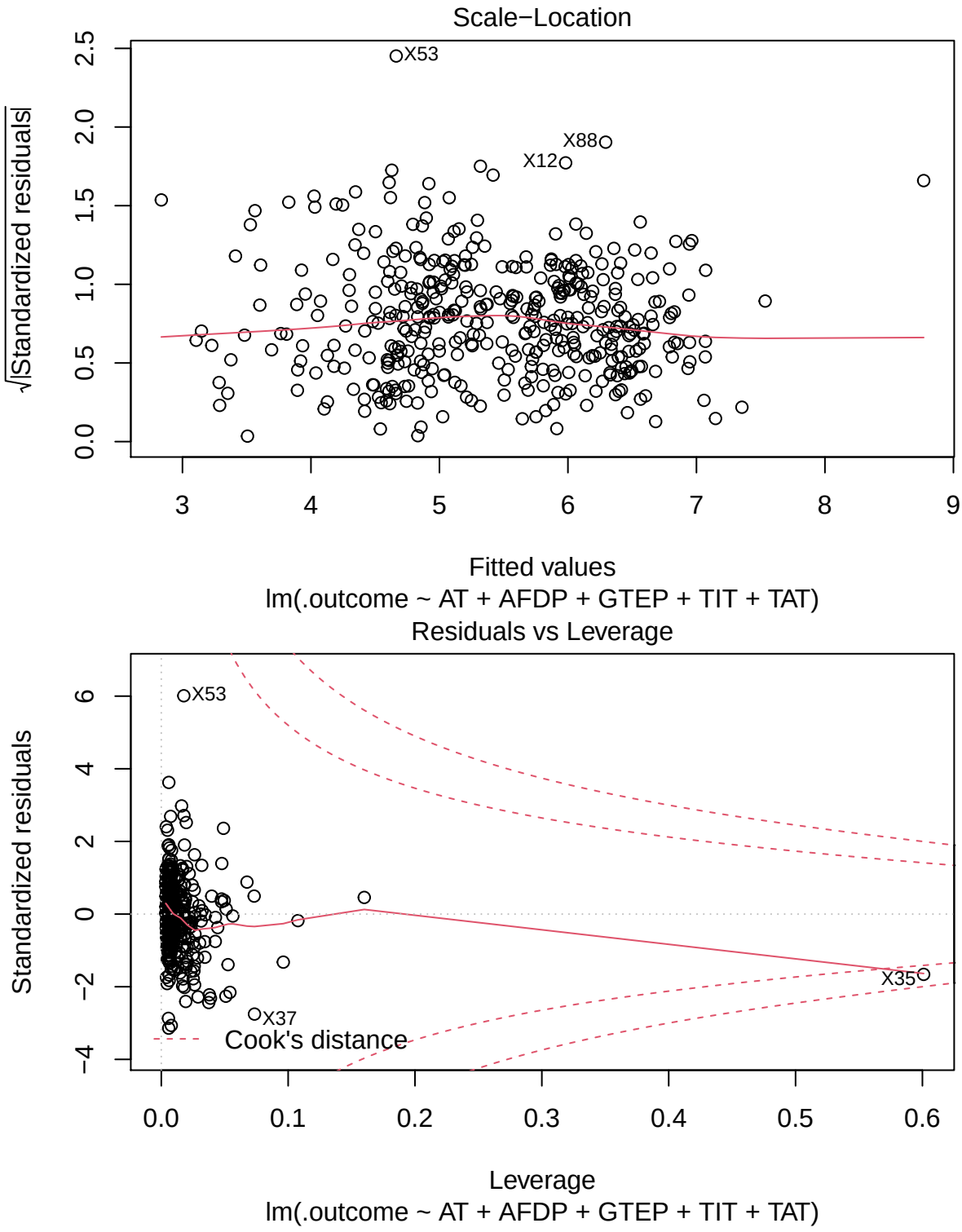

Residuals vs Fitted
Residuals
Fitted values
lm(.outcome ~ AT + AFDP + GTEP + TIT + TAT)
X53
X261
X11
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(.outcome ~ AT + AFDP + GTEP + TIT + TAT)



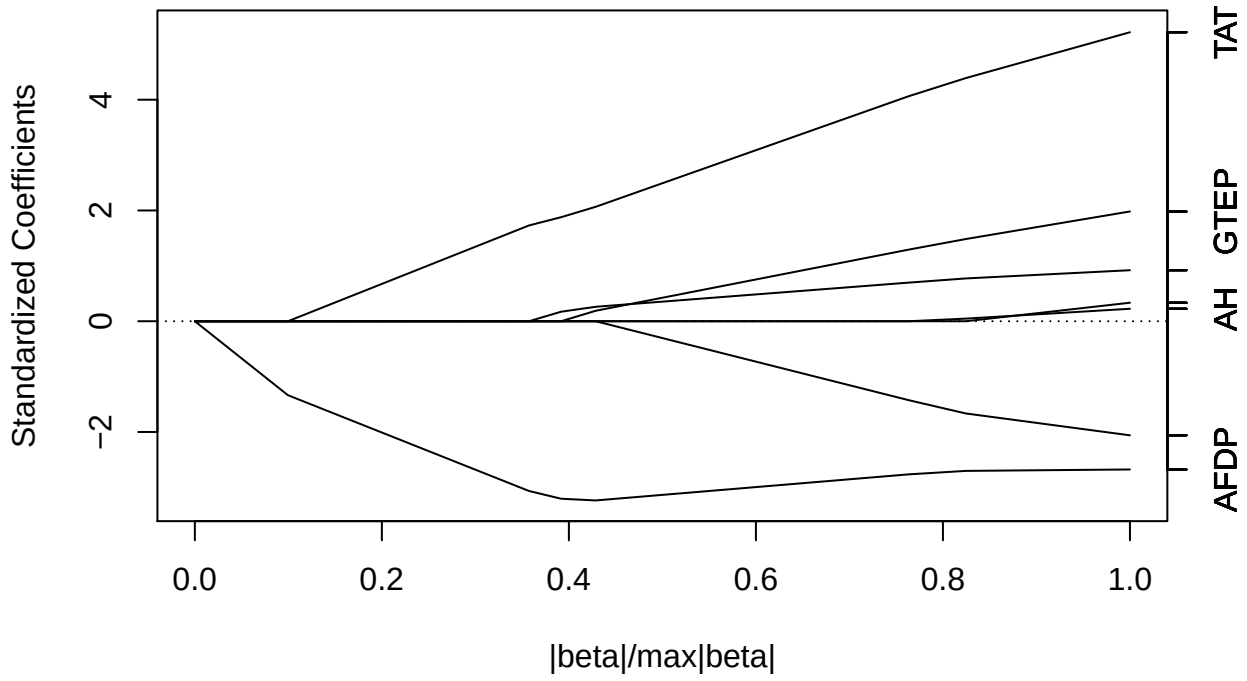



```
plot(high_bc_linear_mod$finalModel)
```


Residuals vs Fitted
Residuals
Fitted values
lm(.outcome ~ AT + AFDP + GTEP + TIT + TAT)
X53
X88
X12
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(.outcome ~ AT + AFDP + GTEP + TIT + TAT)
X53
X88
X12




```
plot(high_lasso_mod$finalModel)
```



#Outlier Search

```
# plot(cooks.distance(all_linear_mod$finalModel))
# plot(cooks.distance(typical_linear_mod$finalModel))
# plot(cooks.distance(high_linear_mod$finalModel))
#
# which(cooks.distance(all_linear_mod$finalModel) > .02)
# which(cooks.distance(typical_linear_mod$finalModel) > .1)
# which(cooks.distance(high_linear_mod$finalModel) > .2)



# plot(cooks.distance(all_bc_linear_mod$finalModel))
# plot(cooks.distance(typical_bc_linear_mod$finalModel))
# plot(cooks.distance(high_bc_linear_mod$finalModel))
#
# which(cooks.distance(all_bc_linear_mod$finalModel) > .01)
# which(cooks.distance(typical_bc_linear_mod$finalModel) > .1)
# which(cooks.distance(high_bc_linear_mod$finalModel) > .2)
```

Decision Trees

#All Data

```
# install.packages('tree')
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v purrr   0.3.4     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
## x dplyr::select() masks MASS::select()
```

```
library(tree)
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

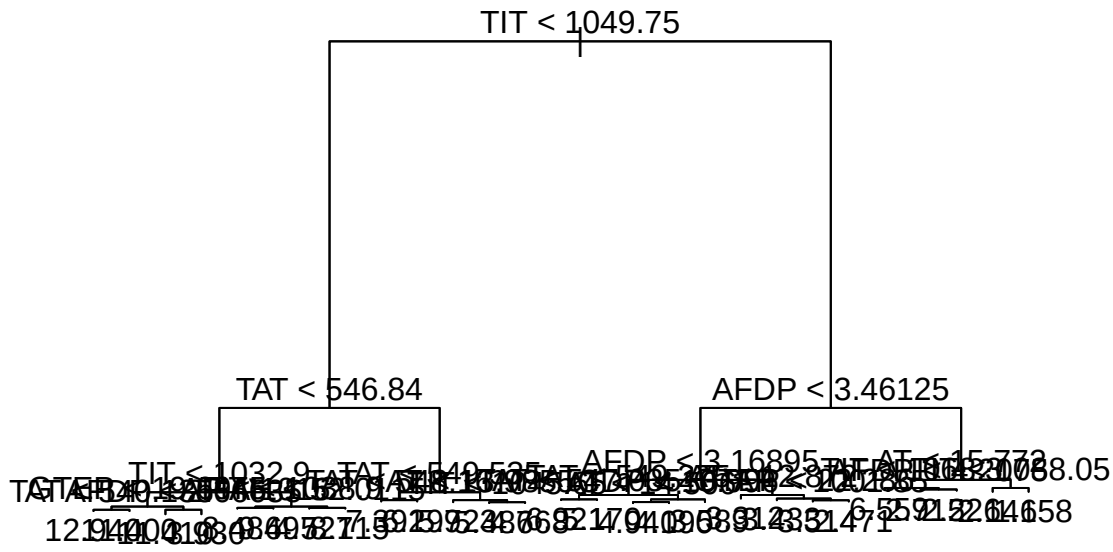
```
RMSE <- function(y, y_hat) {
  rmse <- sqrt(sum(((y_hat - y)^2)/length(y)))
  print(rmse)
}
```

```
set.seed(10)
train <- gt_2015 %>% dplyr::select(-NOX) %>% sample_frac(0.8)
test <- gt_2015 %>% dplyr::select(-NOX) %>% setdiff(train)
```

```
tree_CO <- tree(CO ~ . , train,
                  control = tree.control(nobs = length(train$CO),
                                         minsize = 4, mindev=0.001), method = "recursive.partition")
summary(tree_CO)
```

```
##
## Regression tree:
## tree(formula = CO ~ ., data = train, control = tree.control(nobs = length(train$CO),
##     minsize = 4, mindev = 0.001), method = "recursive.partition")
## Variables actually used in tree construction:
## [1] "TIT"  "TAT"  "GTEP" "AFDP" "AP"   "AH"   "AT"   "CDP"
## Number of terminal nodes:  27
## Residual mean deviance:  0.5043 = 2958 / 5867
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -6.20200 -0.33650 -0.03364  0.00000  0.28310 10.19000
```

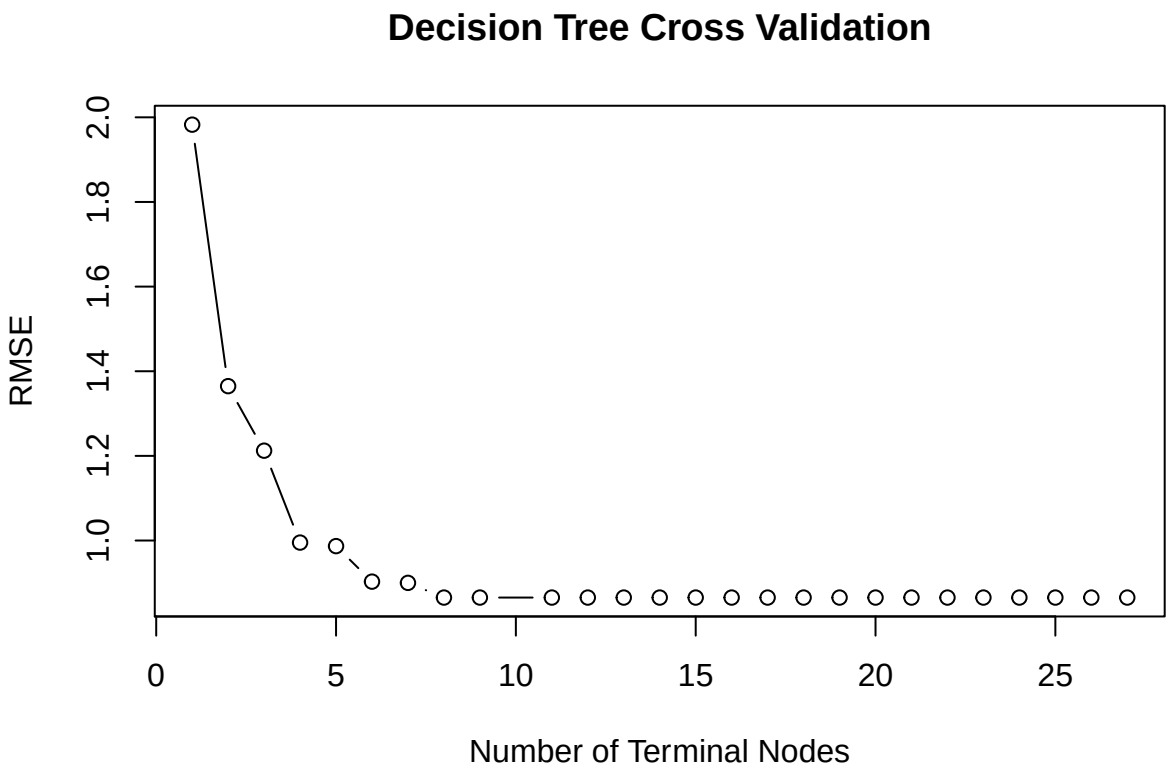
```
plot(tree_CO)
text(tree_CO, pretty = 0)
```




```
tree_pred <- predict(tree_CO, test)
RMSE(test$CO, tree_pred)
```

```
## [1] 0.8184892
```

```
cv_info <- cv.tree(tree_CO, FUN = prune.tree)
plot(cv_info$size, sqrt(cv_info$dev / nrow(train)), type = "b", xlab = "Number of Terminal Nodes", ylab
```

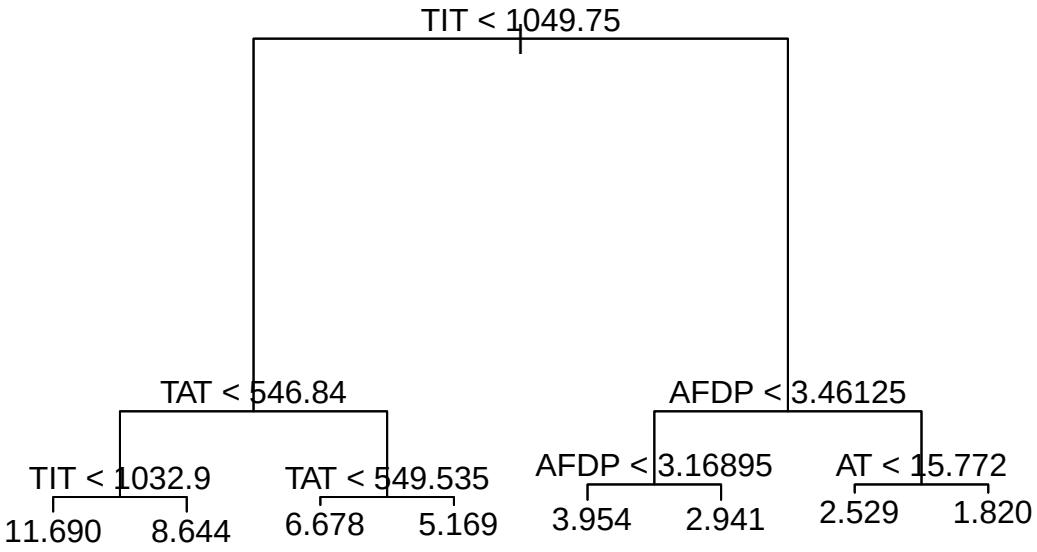



```
pruned_tree <- prune.tree(tree_CO, best = 8)
summary(pruned_tree)
```

```
##
## Regression tree:
```

```
## snip.tree(tree = tree_CO, nodes = c(9L, 10L, 14L, 11L, 13L, 15L,
## 8L, 12L))
## Variables actually used in tree construction:
## [1] "TIT"  "TAT"  "AFDP" "AT"
## Number of terminal nodes:  8
## Residual mean deviance:  0.7172 = 4222 / 5886
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -8.76500 -0.43270 -0.04481  0.00000  0.35110 10.02000
```

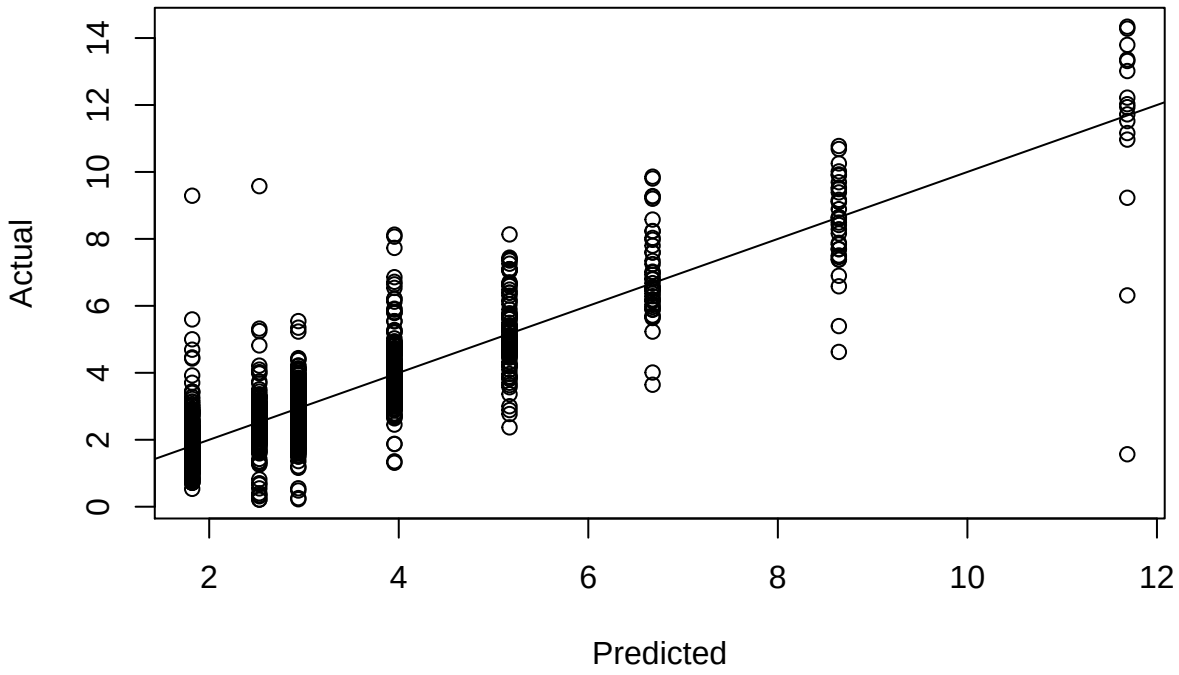
```
plot(pruned_tree)
text(pruned_tree, pretty = 0)
```




```
tree_pred <- predict(pruned_tree, test)
RMSE(test$CO, tree_pred)
```

```
## [1] 0.9111945
```

```
plot(tree_pred, test$CO, xlab = "Predicted", ylab = "Actual")
abline(0, 1)
```



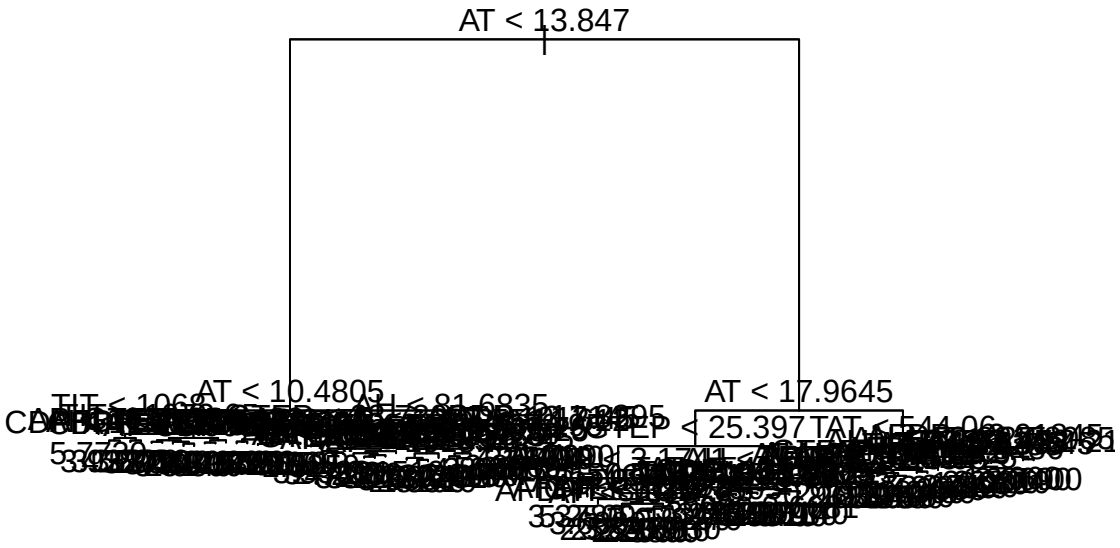
#Typical Energy Yield (127-134.5)

```
set.seed(10)
train_typical <- gt_2015_typical %>% dplyr::select(-NOX) %>% sample_frac(0.8)
test_typical <- gt_2015_typical %>% dplyr::select(-NOX) %>% setdiff(train_typical)
```

```
tree_CO_typical <- tree(CO ~ . , train_typical,
                        control = tree.control(nobs = length(train_typical$CO),
                                               minsize = 4, mindev=0.001), method = "recursive.partition")
summary(tree_CO_typical)
```

```
##
## Regression tree:
## tree(formula = CO ~ ., data = train_typical, control = tree.control(nobs = length(train_typical$CO),
##     minsize = 4, mindev = 0.001), method = "recursive.partition")
## Number of terminal nodes:  89
## Residual mean deviance:  0.1595 = 267 / 1674
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.60600 -0.23060 -0.01263  0.00000  0.21300  2.60600
```

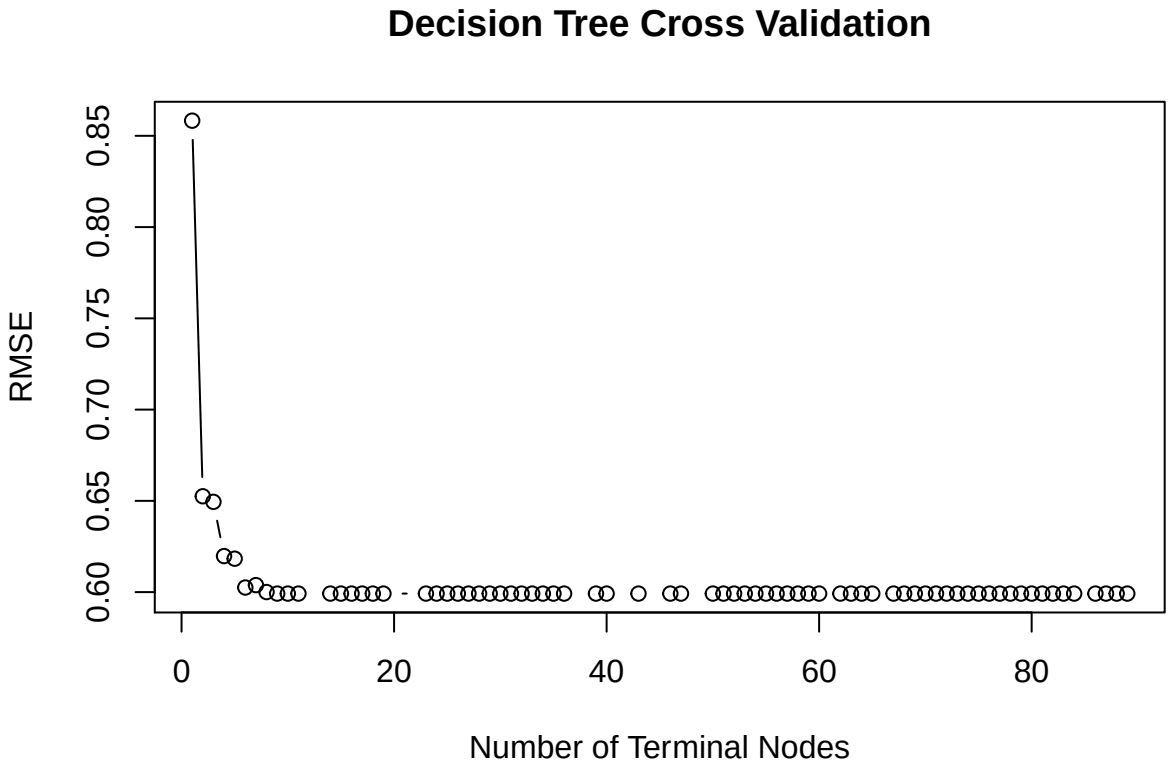
```
plot(tree_CO_typical)
text(tree_CO_typical, pretty = 0)
```




```
tree_pred_typical <- predict(tree_CO_typical, test_typical)
RMSE(test_typical$CO, tree_pred_typical)
```

```
## [1] 0.6649702
```

```
cv_info_typical <- cv.tree(tree_CO_typical, FUN = prune.tree)
plot(cv_info_typical$size, sqrt(cv_info_typical$dev / nrow(train_typical)), type = "b", xlab = "Number
```

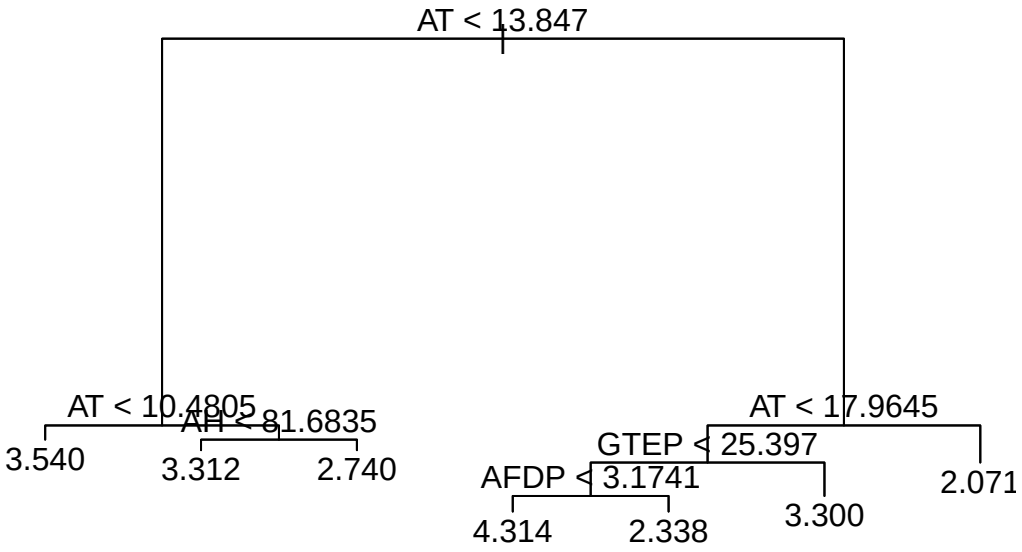



```
pruned_tree_typical <- prune.tree(tree_CO_typical, best = 7)
summary(pruned_tree_typical)
```

```
##
## Regression tree:
```

```
## snip.tree(tree = tree_CO_typical, nodes = c(11L, 24L, 10L, 25L,
## 7L, 13L, 4L))
## Variables actually used in tree construction:
## [1] "AT"   "AH"   "GTEP" "AFDP"
## Number of terminal nodes:  7
## Residual mean deviance:  0.3205 = 562.7 / 1756
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.04300 -0.28570 -0.02228  0.00000  0.24330  5.51500
```

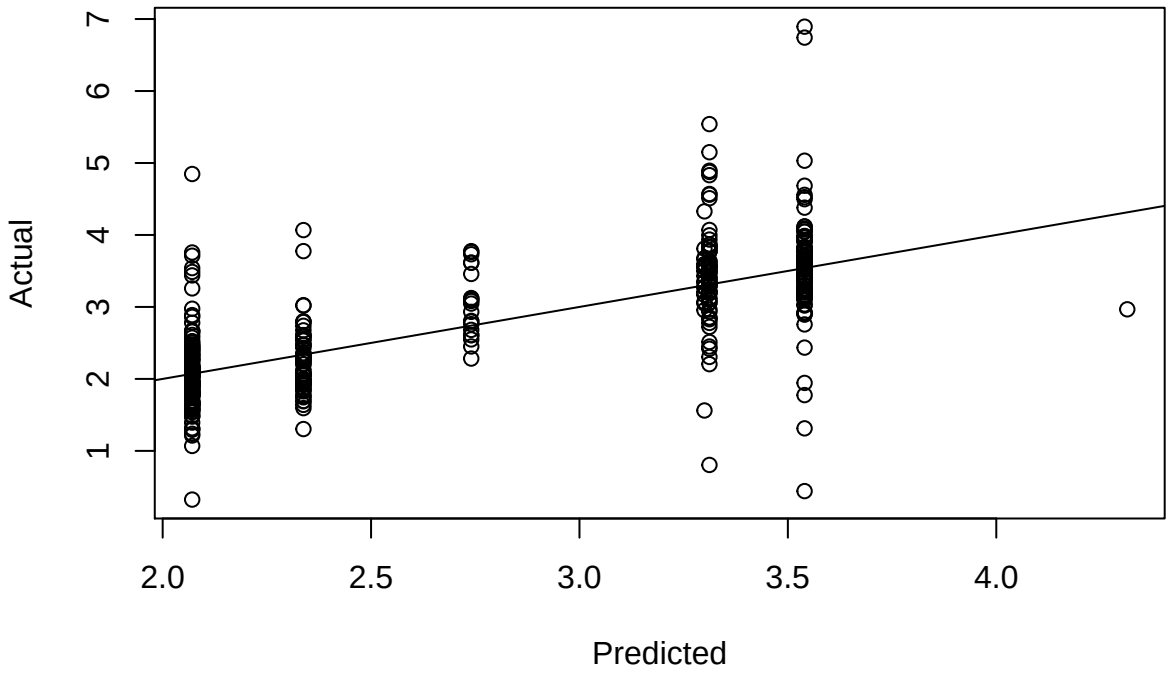
```
plot(pruned_tree_typical)
text(pruned_tree_typical, pretty = 0)
```




```
tree_pred_typical <- predict(pruned_tree_typical, test_typical)
RMSE(test_typical$CO, tree_pred_typical)
```

```
## [1] 0.615961
```

```
plot(tree_pred_typical, test_typical$CO, xlab = "Predicted", ylab = "Actual")
abline(0, 1)
```



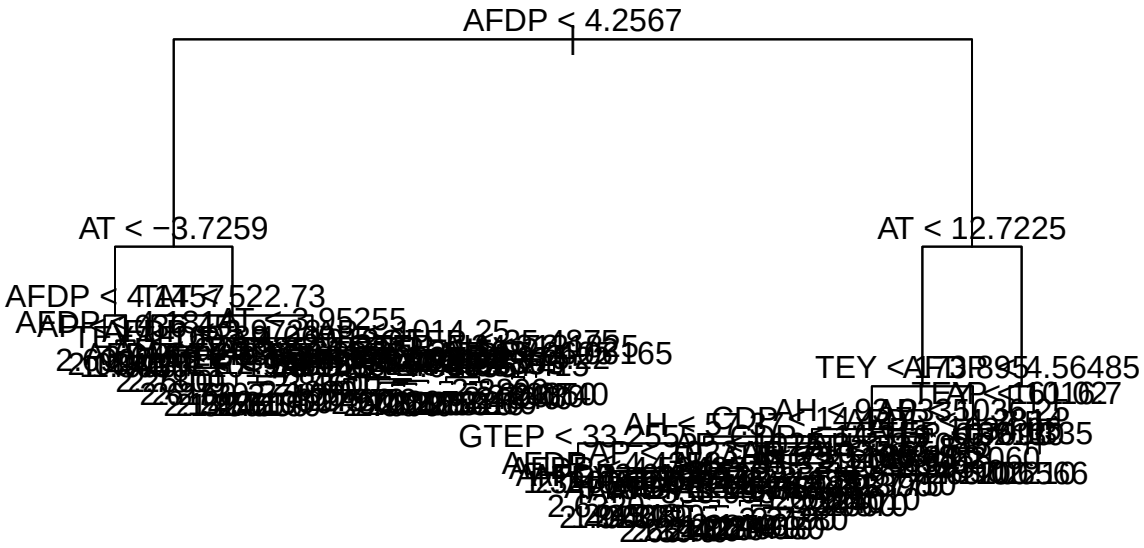
#High Energy Yield (160+)

```
set.seed(10)
train_high <- gt_2015_high %>% dplyr::select(-NOX) %>% sample_frac(0.8)
test_high <- gt_2015_high %>% dplyr::select(-NOX) %>% setdiff(train_high)
```

```
tree_CO_high <- tree(CO ~ . , train_high,
                     control = tree.control(nobs = length(train_high$CO),
                                            minsize = 4, mindev=0.001), method = "recursive.partition")
summary(tree_CO_high)
```

```
##
## Regression tree:
## tree(formula = CO ~ ., data = train_high, control = tree.control(nobs = length(train_high$CO),
##     minsize = 4, mindev = 0.001), method = "recursive.partition")
## Number of terminal nodes:  79
## Residual mean deviance:  0.03494 = 8.909 / 255
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.57980 -0.08404  0.00617  0.00000  0.09731  0.65880
```

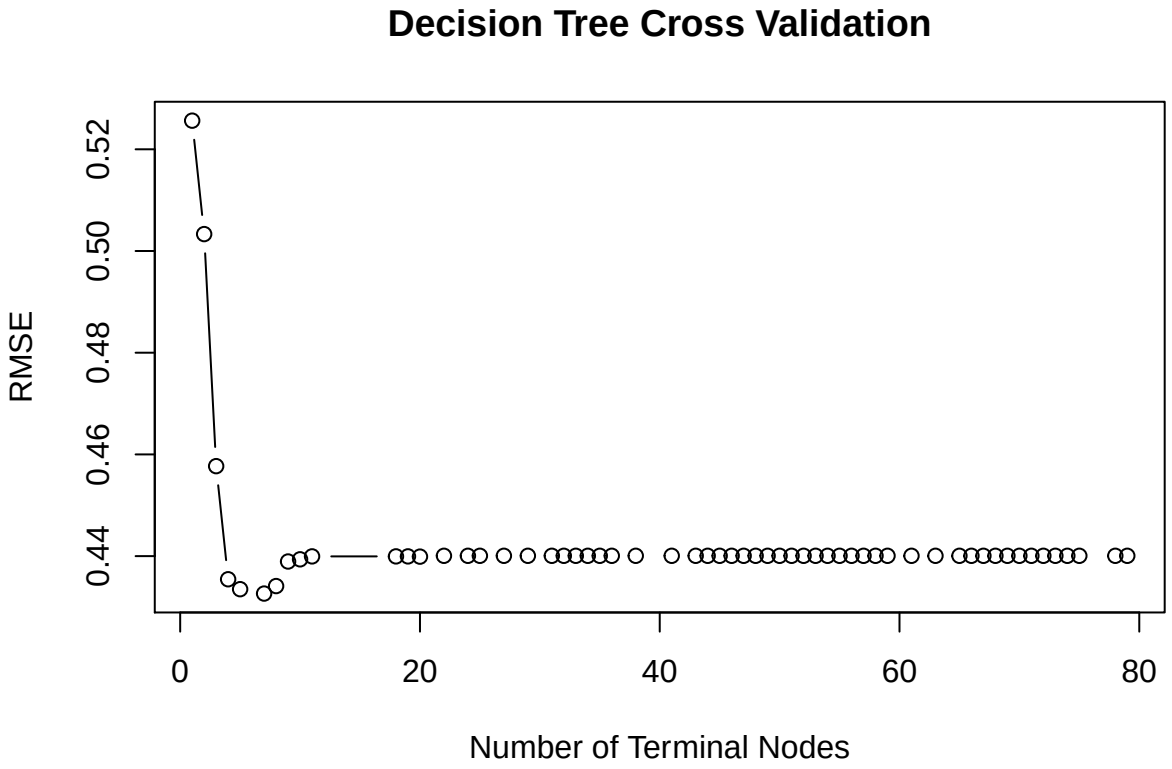
```
plot(tree_CO_high)
text(tree_CO_high, pretty = 0)
```




```
tree_pred_high <- predict(tree_CO_high, test_high)
RMSE(test_high$CO, tree_pred_high)
```

```
## [1] 0.4877416
```

```
cv_info_high <- cv.tree(tree_CO_high, FUN = prune.tree)
plot(cv_info_high$size, sqrt(cv_info_high$dev / nrow(train_high)), type = "b", xlab = "Number of Termina
```

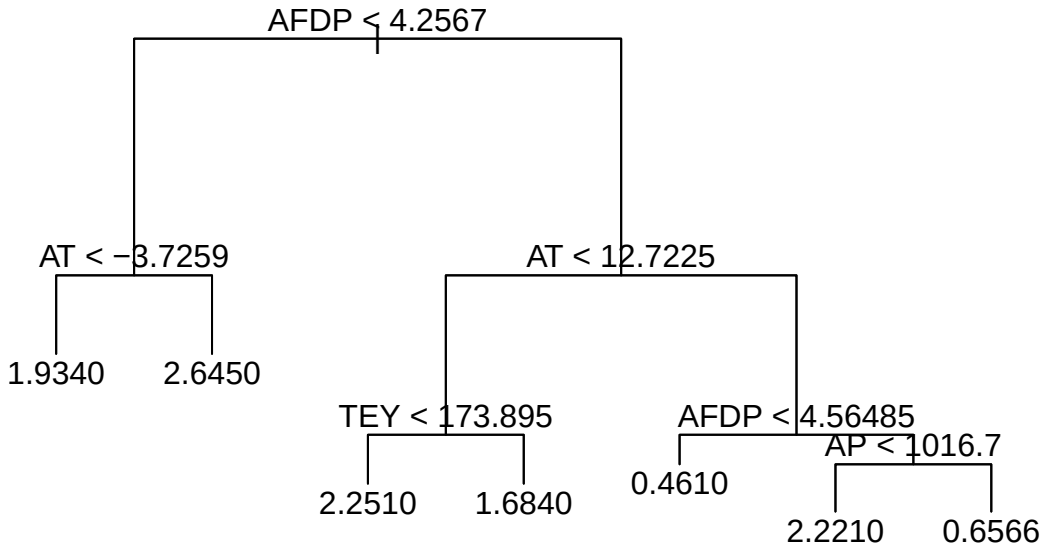



```
pruned_tree_high <- prune.tree(tree_CO_high, best = 6)
summary(pruned_tree_high)
```

```
##
## Regression tree:
```

```
## snip.tree(tree = tree_CO_high, nodes = c(14L, 12L, 13L, 5L, 4L
## ))
## Variables actually used in tree construction:
## [1] "AFDP" "AT"   "TEY"  "AP"
## Number of terminal nodes:  7
## Residual mean deviance:  0.127 = 41.53 / 327
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.76800 -0.19440 -0.01148  0.00000  0.21000  1.45000
```

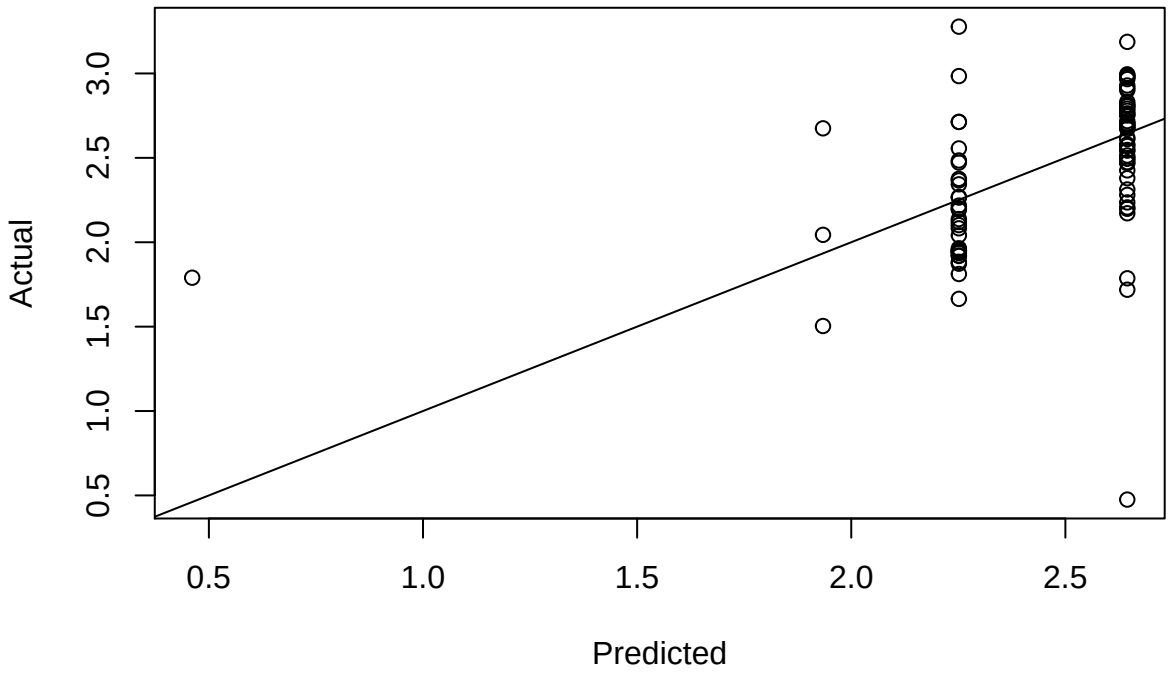
```
plot(pruned_tree_high)
text(pruned_tree_high, pretty = 0)
```




```
tree_pred_high <- predict(pruned_tree_high, test_high)
RMSE(test_high$CO, tree_pred_high)
```

```
## [1] 0.427759
```

```
plot(tree_pred_high, test_high$CO, xlab = "Predicted", ylab = "Actual")
abline(0, 1)
```




```
library(knitr)
all_linear_rmse     <- all_linear_mod$results$RMSE
typical_linear_rmse <- typical_linear_mod$results$RMSE
high_linear_rmse    <- high_linear_mod$results$RMSE

all_lass_rmse     <- max(all_lasso_mod$results$RMSE)
typical_lass_rmse <- max(typical_lasso_mod$results$RMSE)
high_lass_rmse    <- max(high_lasso_mod$results$RMSE)

all_tree_rmse     <- RMSE(test$CO, tree_pred)
```

```
## [1] 0.9111945
```

```
typical_tree_rmse <- RMSE(test_typical$CO, tree_pred_typical)
```

```
## [1] 0.615961
```

```
high_tree_rmse    <- RMSE(test_high$CO, tree_pred_high)
```

```
## [1] 0.427759
```

```
bc_all_rmse     <- all_bc_linear_mod$results$RMSE
bc_typical_rmse <- typical_bc_linear_mod$results$RMSE
bc_high_rmse    <- high_bc_linear_mod$results$RMSE



RMSE_Table <- matrix(c(round(all_linear_mod$results$RMSE        ,digits = 4),
                       round(typical_linear_mod$results$RMSE    ,digits = 4),
                       round(high_linear_mod$results$RMSE       ,digits = 4),
                       round(all_bc_linear_mod$results$RMSE     ,digits = 4),
                       round(typical_bc_linear_mod$results$RMSE ,digits = 4),
                       round(high_bc_linear_mod$results$RMSE    ,digits = 4),
                       round(max(all_lasso_mod$results$RMSE)    ,digits = 4),
                       round(max(typical_lasso_mod$results$RMSE),digits = 4),
```

```
                      round(max(high_lasso_mod$results$RMSE)   ,digits = 4),
                      round(all_tree_rmse                      ,digits = 4),
                      round(typical_tree_rmse                  ,digits = 4),
                      round(high_tree_rmse                     ,digits = 4 ))
                    ,ncol=3, byrow=TRUE)
colnames(RMSE_Table) <- c("Overall Production Range", "Typical Production Range (127-134.5)", "High Pro
rownames(RMSE_Table) <- c("Linear Regression", "Box-Cox Transformed", "Lasso", "Decision Tree" )
RMSE_Table <- as.table(RMSE_Table)
RMSE_Table <- kable(RMSE_Table)
RMSE_Table
```

	Overall Production Range	Typical Production Range (127-134.5)	High Production Range (160+)
Linear Regression	1.1088	0.6293	0.4428
Box-Cox Transformed	0.0756	0.2528	1.6889
Lasso	1.7271	0.7939	0.4772
Decision Tree	0.9112	0.6160	0.4278