SS3859 Group Project

Rui Zhu

Calculate p-value for each predictor

```
df <- read.csv(file = 'PRSA_data_2010.1.1-2014.12.31.csv')
df=na.omit(df)
set.seed(3859)
index <- sample(1:nrow(df), 2000)
df=df[index, ]
nrow(df)
```

```
## [1] 2000
```

```
head(df)
```

```
##          No year month day hour pm2.5 DEWP TEMP PRES cbwd    Iws Is Ir
## 12224 12224 2011     5  25    7    66   12   20 1014   SE 185.07  0  0
## 33448 33448 2013    10  25   15    25   -7   19 1025   NW   1.79  0  0
## 13028 13028 2011     6  27   19    53   15   28  999   SE  22.35  0  0
## 31240 31240 2013     7  25   15    30   15   36 1000   SE   7.15  0  0
## 14633 14633 2011     9   2   16    57   18   30 1008   SE   4.02  0  0
## 15579 15579 2011    10  12    2   221   13   14 1020   NE   1.78  0  0
```

```
model1=lm(pm2.5 ~ DEWP+TEMP+PRES+cbwd+Iws+Is+Ir, data=df)
summary(model1)
```

```
##
## Call:
## lm(formula = pm2.5 ~ DEWP + TEMP + PRES + cbwd + Iws + Is + Ir,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -161.16  -52.60  -14.04   33.52  424.90
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1663.80214  332.84996   4.999 6.28e-07 ***
## DEWP           3.70727    0.24686  15.018  < 2e-16 ***
## TEMP          -6.22270    0.31006 -20.069  < 2e-16 ***
## PRES          -1.45221    0.32551  -4.461 8.60e-06 ***
## cbwdNE       -27.81937    6.52102  -4.266 2.08e-05 ***
## cbwdNW       -29.17477    5.36890  -5.434 6.19e-08 ***
## cbwdSE         6.02305    5.04614   1.194    0.233
## Iws           -0.29018    0.04062  -7.144 1.27e-12 ***
## Is            -2.00165    1.74745  -1.145    0.252
## Ir            -6.06433    1.25242  -4.842 1.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.75 on 1990 degrees of freedom
## Multiple R-squared:  0.2602, Adjusted R-squared:  0.2568
## F-statistic: 77.76 on 9 and 1990 DF,  p-value: < 2.2e-16
```

Some hypothese that which predictors are not significant

- By observing the summary table of the full model, we make the null hypothesis that the Is is not significantly important to explain this model.

```
#reduced model without Is
model2=lm(pm2.5 ~ DEWP+TEMP+PRES+Iws+cbwd+Ir, data=df)
summary(model2)

##
## Call:
## lm(formula = pm2.5 ~ DEWP + TEMP + PRES + Iws + cbwd + Ir, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -160.70  -52.56  -14.16   33.80  425.48
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1661.20329  332.86831   4.991 6.54e-07 ***
## DEWP           3.68358    0.24601  14.973  < 2e-16 ***
## TEMP          -6.17968    0.30780 -20.077  < 2e-16 ***
## PRES          -1.45020    0.32554  -4.455 8.86e-06 ***
## Iws           -0.29319    0.04054  -7.232 6.74e-13 ***
## cbwdNE       -27.69136    6.52057  -4.247 2.27e-05 ***
## cbwdNW       -29.05419    5.36828  -5.412 6.98e-08 ***
## cbwdSE         5.77988    5.04207   1.146    0.252
## Ir            -6.03967    1.25233  -4.823 1.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.76 on 1991 degrees of freedom
## Multiple R-squared:  0.2597, Adjusted R-squared:  0.2567
## F-statistic:  87.3 on 8 and 1991 DF,  p-value: < 2.2e-16

anova(model1,model2)

## Analysis of Variance Table
##
## Model 1: pm2.5 ~ DEWP + TEMP + PRES + cbwd + Iws + Is + Ir
## Model 2: pm2.5 ~ DEWP + TEMP + PRES + Iws + cbwd + Ir
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1   1990 12656854
## 2   1991 12665200 -1   -8345.3 1.3121 0.2522
```

- The large f statistics value in anova indicates there is no significant difference between the two models. Therefore, we fail to reject the null hypothesis that Is is significantly important to explain the model.

Test for interactions(2 way interaction)

```
model4=lm(pm2.5 ~
DEWP+TEMP+PRES+cbwd+Iws+Ir+Is+I(DEWP*TEMP)+I(DEWP*PRES)+I(DEWP*Iws)+I(DEWP*Ir
)+I(TEMP*PRES)+I(TEMP*Iws)+I(TEMP*Ir)+I(PRES*Iws)+I(PRES*Ir)+I(Iws*Ir)+I(Is*T
EMP)+I(Is*DEWP)+I(Is*PRES)+I(Is*Iws)+I(Is*Ir), data=df)


summary(model4)

##
## Call:
## lm(formula = pm2.5 ~ DEWP + TEMP + PRES + cbwd + Iws + Ir + Is +
##     I(DEWP * TEMP) + I(DEWP * PRES) + I(DEWP * Iws) + I(DEWP *
##     Ir) + I(TEMP * PRES) + I(TEMP * Iws) + I(TEMP * Ir) + I(PRES *
##     Iws) + I(PRES * Ir) + I(Iws * Ir) + I(Is * TEMP) + I(Is *
##     DEWP) + I(Is * PRES) + I(Is * Iws) + I(Is * Ir), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -179.68  -49.14  -11.57   32.09  406.91
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.440e+03  5.057e+02   2.847 0.004460 **
## DEWP           -1.639e+02  2.893e+01  -5.664 1.70e-08 ***
## TEMP            6.369e+01  2.922e+01   2.179 0.029433 *
## PRES           -1.211e+00  4.960e-01  -2.441 0.014720 *
## cbwdNE         -2.457e+01  6.371e+00  -3.857 0.000118 ***
## cbwdNW         -2.643e+01  5.278e+00  -5.007 6.02e-07 ***
## cbwdSE          1.139e+01  5.069e+00   2.246 0.024789 *
## Iws            -2.183e+01  7.580e+00  -2.880 0.004020 **
## Ir              1.306e+02  2.807e+02   0.465 0.641719
## Is              9.093e+02  1.607e+03   0.566 0.571595
## I(DEWP * TEMP) -1.144e-02  2.414e-02  -0.474 0.635621
## I(DEWP * PRES)  1.656e-01  2.830e-02   5.852 5.67e-09 ***
## I(DEWP * Iws)  -2.519e-02  5.697e-03  -4.422 1.03e-05 ***
## I(DEWP * Ir)    3.998e-01  1.228e+00   0.325 0.744845
## I(TEMP * PRES) -6.968e-02  2.876e-02  -2.423 0.015493 *
## I(TEMP * Iws)   3.868e-02  7.010e-03   5.518 3.88e-08 ***
## I(TEMP * Ir)   -3.342e-01  1.208e+00  -0.277 0.782140
## I(PRES * Iws)   2.061e-02  7.412e-03   2.781 0.005467 **
## I(PRES * Ir)   -1.373e-01  2.737e-01  -0.502 0.616012
## I(Iws * Ir)     5.199e-02  3.980e-02   1.306 0.191637
## I(Is * TEMP)    2.407e+00  3.878e+00   0.621 0.534827
## I(Is * DEWP)   -1.439e+00  3.504e+00  -0.411 0.681402
## I(Is * PRES)   -8.956e-01  1.564e+00  -0.573 0.566965
## I(Is * Iws)     8.735e-02  9.275e-02   0.942 0.346407
## I(Is * Ir)             NA         NA      NA       NA
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.48 on 1976 degrees of freedom
## Multiple R-squared:  0.3067, Adjusted R-squared:  0.2986
## F-statistic: 38.01 on 23 and 1976 DF,  p-value: < 2.2e-16
```

- By observing the p-value of each predictior, we make the null hypothesis that DEWP * TEMP, DEWP * Ir, TEMP * Ir and PRES * Ir, Is * TEMP, Is * DEWP and Is * PRES are not significantly important to explain this model.

```
#reduced model without DEWP * TEMP, DEWP * Ir, TEMP * Ir and PRES * Ir, Is * TEMP, Is * DEWP and Is * PRES
model5=lm(pm2.5 ~ DEWP+TEMP+PRES+cbwd+Iws+Ir+Is+I(DEWP*PRES)+I(DEWP*Iws)+I(TEMP*PRES)+I(TEMP*Iws)+I(PRES*Iws)+I(Iws*Ir)+I(Is*Iws)+I(Is*Ir), data=df)
summary(model5)

##
## Call:
## lm(formula = pm2.5 ~ DEWP + TEMP + PRES + cbwd + Iws + Ir + Is +
##     I(DEWP * PRES) + I(DEWP * Iws) + I(TEMP * PRES) + I(TEMP *
##     Iws) + I(PRES * Iws) + I(Iws * Ir) + I(Is * Iws) + I(Is *
##     Ir), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -178.60  -49.25  -12.19   32.71  405.14
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.588e+03  4.762e+02   3.335 0.000869 ***
## DEWP           -1.663e+02  2.445e+01  -6.804 1.35e-11 ***
## TEMP            5.604e+01  2.741e+01   2.044 0.041064 *
## PRES           -1.357e+00  4.667e-01  -2.909 0.003672 **
## cbwdNE         -2.502e+01  6.357e+00  -3.936 8.56e-05 ***
## cbwdNW         -2.639e+01  5.260e+00  -5.016 5.74e-07 ***
## cbwdSE          1.081e+01  5.043e+00   2.144 0.032180 *
## Iws            -2.304e+01  7.341e+00  -3.138 0.001724 **
## Ir             -7.215e+00  1.562e+00  -4.620 4.08e-06 ***
## Is             -1.221e+01  3.702e+00  -3.299 0.000989 ***
## I(DEWP * PRES)  1.678e-01  2.406e-02   6.977 4.10e-12 ***
## I(DEWP * Iws)  -2.480e-02  5.647e-03  -4.391 1.19e-05 ***
## I(TEMP * PRES) -6.216e-02  2.698e-02  -2.304 0.021320 *
## I(TEMP * Iws)   3.980e-02  6.868e-03   5.794 7.95e-09 ***
## I(PRES * Iws)   2.179e-02  7.178e-03   3.036 0.002430 **
## I(Iws * Ir)     4.407e-02  1.827e-02   2.412 0.015956 *
## I(Is * Iws)     1.431e-01  4.407e-02   3.247 0.001187 **
## I(Is * Ir)             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 77.44 on 1983 degrees of freedom
## Multiple R-squared:  0.305,  Adjusted R-squared:  0.2994
## F-statistic: 54.38 on 16 and 1983 DF,  p-value: < 2.2e-16
```

```
anova(model4,model5)
```

```
## Analysis of Variance Table
##
## Model 1: pm2.5 ~ DEWP + TEMP + PRES + cbwd + Iws + Ir + Is + I(DEWP *
##     TEMP) + I(DEWP * PRES) + I(DEWP * Iws) + I(DEWP * Ir) + I(TEMP *
##     PRES) + I(TEMP * Iws) + I(TEMP * Ir) + I(PRES * Iws) + I(PRES *
##     Ir) + I(Iws * Ir) + I(Is * TEMP) + I(Is * DEWP) + I(Is *
##     PRES) + I(Is * Iws) + I(Is * Ir)
## Model 2: pm2.5 ~ DEWP + TEMP + PRES + cbwd + Iws + Ir + Is + I(DEWP *
##     PRES) + I(DEWP * Iws) + I(TEMP * PRES) + I(TEMP * Iws) +
##     I(PRES * Iws) + I(Iws * Ir) + I(Is * Iws) + I(Is * Ir)
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1   1976 11860707
## 2   1983 11890475 -7    -29768 0.7085 0.6649
```

- The large p-value in anova indicates there is no significant difference between the two models.

Variable selection (which variables to keep, based on previous results and AIC, BIC or PRESS test)

```
nullfit <- lm(pm2.5~1,data=df)
stepAppro_aic = step(nullfit,
                     scope = pm2.5 ~
DEWP+TEMP+PRES+cbwd+Iws+Ir+Is+I(DEWP*TEMP)+I(DEWP*PRES)+I(DEWP*Iws)+I(DEWP*Ir
)+I(TEMP*PRES)+I(TEMP*Iws)+I(TEMP*Ir)+I(PRES*Iws)+I(PRES*Ir)+I(Iws*Ir)+I(Is*T
EMP)+I(Is*DEWP)+I(Is*PRES)+I(Is*Iws)+I(Is*Ir),
                     direction = "forward",
                     trace = 0)

stepAppro_bic <- step(model4,
                      direction = "backward",
                      k=log(nrow(df)),
                      trace=FALSE)
stepAppro_aic
```

```
##
## Call:
## lm(formula = pm2.5 ~ Iws + I(TEMP * PRES) + I(DEWP * PRES) +
##     cbwd + DEWP + I(PRES * Ir) + PRES + I(DEWP * Iws) + I(TEMP *
##     Iws) + I(PRES * Iws) + I(Iws * Ir) + I(Is * TEMP) + I(Is *
##     Iws) + I(Is * PRES) + TEMP, data = df)
##
## Coefficients:
##    (Intercept)             Iws  I(TEMP * PRES)  I(DEWP * PRES)
```

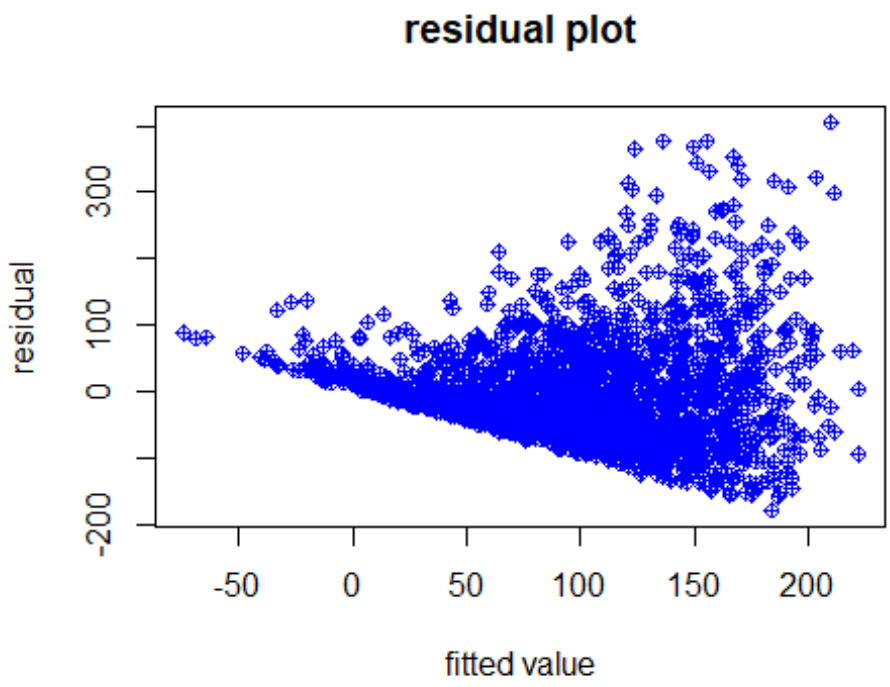
```
cbwdNE
##      1.592e+03      -2.241e+01      -6.215e-02       1.676e-01      -
2.504e+01
##         cbwdNW          cbwdSE            DEWP    I(PRES * Ir)
PRES
##     -2.666e+01       1.101e+01      -1.661e+02      -7.146e-03      -
1.361e+00
##  I(DEWP * Iws)   I(TEMP * Iws)   I(PRES * Iws)     I(Iws * Ir)    I(Is *
TEMP)
##     -2.502e-02       3.932e-02       2.118e-02       4.398e-02
7.758e-01
##    I(Is * Iws)    I(Is * PRES)            TEMP
##      1.371e-01      -9.425e-03       5.602e+01

stepAppro_bic

##
## Call:
## lm(formula = pm2.5 ~ DEWP + PRES + cbwd + Iws + I(DEWP * PRES) +
##     I(DEWP * Iws) + I(TEMP * PRES) + I(TEMP * Iws) + I(PRES *
##     Ir) + I(Is * PRES) + I(Is * Iws), data = df)
##
## Coefficients:
##    (Intercept)            DEWP            PRES          cbwdNE
cbwdNW
##     1765.35515      -117.36455        -1.53470       -25.06620       -
26.99562
##         cbwdSE             Iws  I(DEWP * PRES)   I(DEWP * Iws)  I(TEMP *
PRES)
##       11.35386        -0.69972         0.11974        -0.02789        -
0.00681
##  I(TEMP * Iws)    I(PRES * Ir)    I(Is * PRES)     I(Is * Iws)
##        0.02672        -0.00537        -0.01122         0.13292

anova(stepAppro_aic,stepAppro_bic)

## Analysis of Variance Table
##
## Model 1: pm2.5 ~ Iws + I(TEMP * PRES) + I(DEWP * PRES) + cbwd + DEWP +
##     I(PRES * Ir) + PRES + I(DEWP * Iws) + I(TEMP * Iws) + I(PRES *
##     Iws) + I(Iws * Ir) + I(Is * TEMP) + I(Is * Iws) + I(Is *
##     PRES) + TEMP
## Model 2: pm2.5 ~ DEWP + PRES + cbwd + Iws + I(DEWP * PRES) + I(DEWP *
##     Iws) + I(TEMP * PRES) + I(TEMP * Iws) + I(PRES * Ir) + I(Is *
##     PRES) + I(Is * Iws)
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1   1982 11874882
## 2   1986 11994378 -4   -119497 4.9862 0.0005334 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(asbio)

## Loading required package: tcltk

#model selected by AIC
press(lm(formula = pm2.5 ~ Iws + I(TEMP * PRES) + I(DEWP * PRES) +
    cbwd + DEWP + I(PRES * Ir) + PRES + I(DEWP * Iws) + I(TEMP *
    Iws) + I(PRES * Iws) + I(Iws * Ir) + I(Is * TEMP) + I(Is *
    Iws) + I(Is * PRES) + TEMP, data = df)
)

## [1] 12018669

#model selected by BIC
press(lm(formula = pm2.5 ~ DEWP + PRES + cbwd + Iws + I(DEWP * PRES) +
    I(DEWP * Iws) + I(TEMP * PRES) + I(TEMP * Iws) + I(PRES *
    Ir) + I(Is * PRES) + I(Is * Iws), data = df))

## [1] 12141580
```

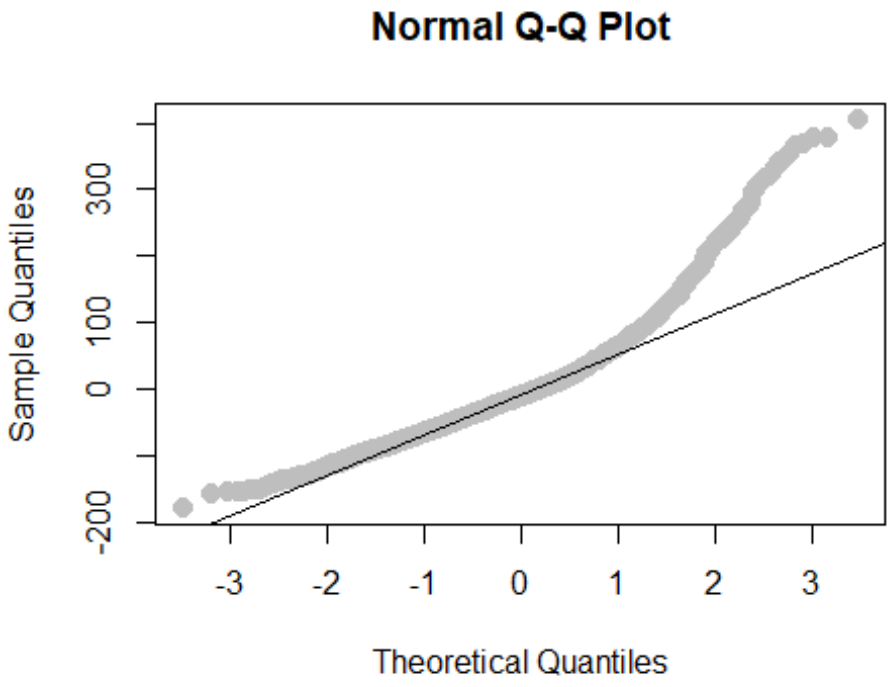
- The PRESS statistic indicates that model selected by AIC is more preferred in this case. However, PRESS might not be approprite although the dataset is reduced already.

Model diagnostics on one well-fit model

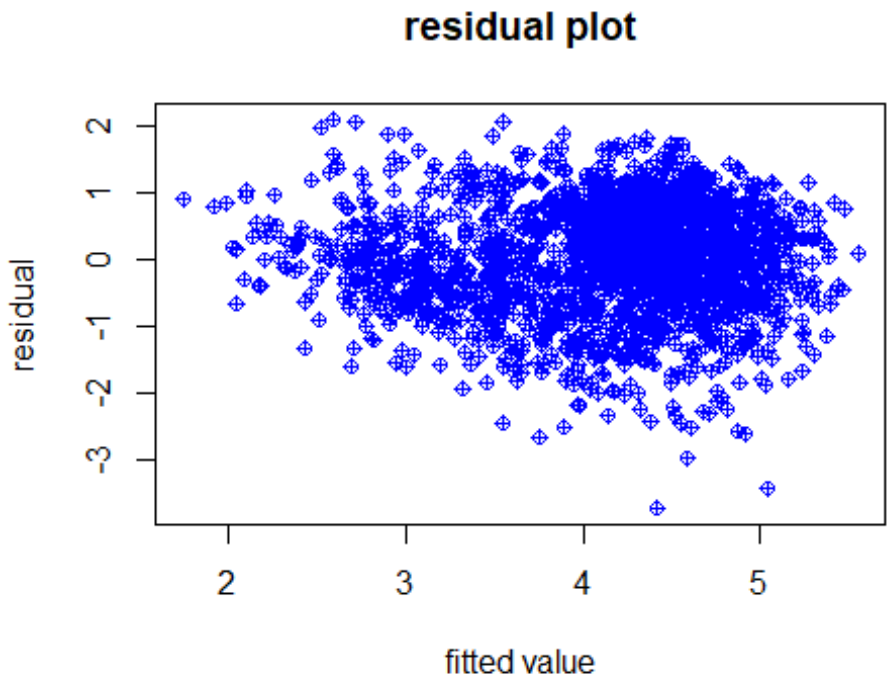
```
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

#we will be using the model selected by AIC in later learning
model=lm(formula = pm2.5 ~ Iws + I(TEMP * PRES) + I(DEWP * PRES) +
    cbwd + DEWP + I(PRES * Ir) + PRES + I(DEWP * Iws) + I(TEMP *
    Iws) + I(PRES * Iws) + I(Iws * Ir) + I(Is * TEMP) + I(Is *
    Iws) + I(Is * PRES) + TEMP, data = df)
plot(fitted(model), resid(model),
     col = "blue", pch = 10,
     xlab = "fitted value",
     ylab = "residual",
     cex=1,
     main = "residual plot")
```




```r
qqnorm(resid(model), col = "grey",pch=20,cex=2)
qqline(resid(model))
```

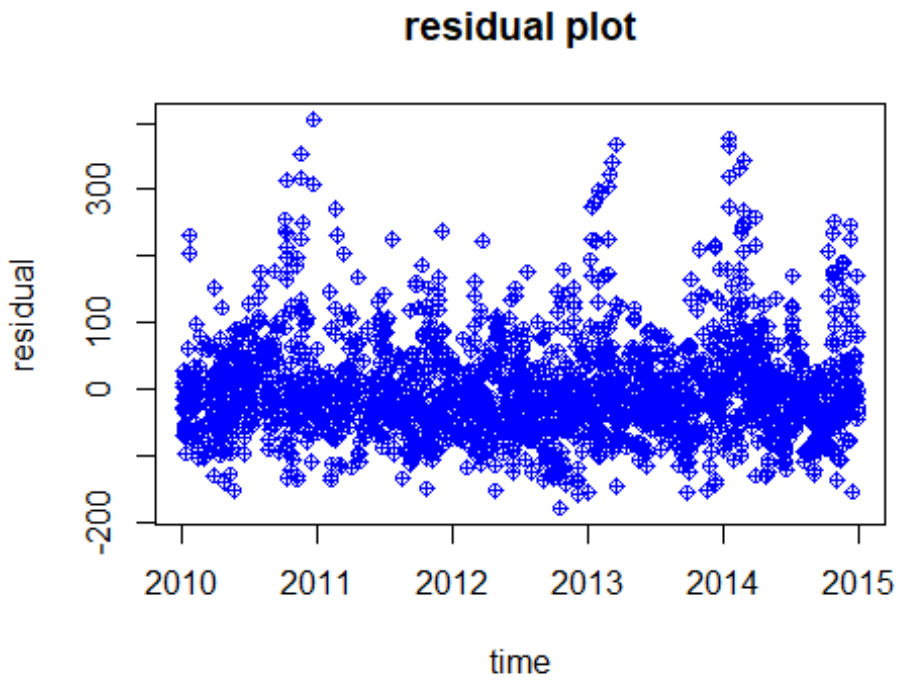


```
loggedModel=lm(formula = log(pm2.5) ~ Iws + I(TEMP * PRES) + I(DEWP * PRES) +
    cbwd + DEWP + I(PRES * Ir) + PRES + I(DEWP * Iws) + I(TEMP *
    Iws) + I(PRES * Iws) + I(Iws * Ir) + TEMP, data = df)
plot(fitted(loggedModel), resid(loggedModel),
     col = "blue", pch = 10,
     xlab = "fitted value",
     ylab = "residual",
     cex=1,
     main = "residual plot")
```




```
bptest(model)

##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 210.02, df = 17, p-value < 2.2e-16

shapiro.test(resid(model))

##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.91623, p-value < 2.2e-16

timeset=c()
residset=c()
```

```
i=1
while (i<nrow(df)){
  timeset=append(timeset,
as.Date(paste(df[i,"month"],df[i,"day"],df[i,"year"],sep="/"), "%m/%d/%Y"))
  residset=append(residset, df[i, "pm2.5"]-predict(model, df[i,]))
  i=i+1
}
plot(timeset, residset,
     col = "blue", pch = 10,
     xlab = "time",
     ylab = "residual",
     cex=1,
     main = "residual plot")
```




```
length(timeset)
```

```
## [1] 1999
```

- Assumptions:
  - Linearity: The residuals distribute systematically and do not exhibit a mean of zero. The linearity assumption is violated.
  - Equal Variance: The small p-value of the BP test indicates that the variance assumption is violated.

- Normality Assumption: The small p-value of SW test indicates that the normality assumption is violated. However, the logged model might hold the normality assumption.
- Independence Assumption: The residual plot against time, the value of random errors are independent. The normality assumption holds.