## Appendix

**Code used to produce plots, tables, and draw conclusions. All written by Edward Huber for purposes of this analysis.**

In [1]:

```
library(tidyr)
library(ggplot2)
library(MASS)
quality = read.csv('airquality.csv')
```

```
Registered S3 methods overwritten by 'ggplot2':
  method         from 
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
```

In [2]:

```
head(quality)
```

| No | year | month | day | hour | pm2.5 | DEWP | TEMP | PRES | cbwd | Iws | Is | Ir |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2010 | 1 | 1 | 0 | NA | -21 | -11 | 1021 | NW | 1.79 | 0 | 0 |
| 2 | 2010 | 1 | 1 | 1 | NA | -21 | -12 | 1020 | NW | 4.92 | 0 | 0 |
| 3 | 2010 | 1 | 1 | 2 | NA | -21 | -11 | 1019 | NW | 6.71 | 0 | 0 |
| 4 | 2010 | 1 | 1 | 3 | NA | -21 | -14 | 1019 | NW | 9.84 | 0 | 0 |
| 5 | 2010 | 1 | 1 | 4 | NA | -20 | -12 | 1018 | NW | 12.97 | 0 | 0 |
| 6 | 2010 | 1 | 1 | 5 | NA | -19 | -10 | 1017 | NW | 16.10 | 0 | 0 |

In [3]:

```
nrow(quality)
```

43824

In [4]:

```
quality2 = drop_na(quality)
row.names(quality2) = 1:nrow(quality2)
names(quality2) = tolower(names(quality2))
nrow(quality2)
```

41757

In [5]:

```
head(quality2)
```

| no | year | month | day | hour | pm2.5 | dewp | temp | pres | cbwd | iws | is | ir |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 2010 | 1 | 2 | 0 | 129 | -16 | -4 | 1020 | SE | 1.79 | 0 | 0 |
| 26 | 2010 | 1 | 2 | 1 | 148 | -15 | -4 | 1020 | SE | 2.68 | 0 | 0 |
| 27 | 2010 | 1 | 2 | 2 | 159 | -11 | -5 | 1021 | SE | 3.57 | 0 | 0 |
| 28 | 2010 | 1 | 2 | 3 | 181 | -7 | -5 | 1022 | SE | 5.36 | 1 | 0 |
| 29 | 2010 | 1 | 2 | 4 | 138 | -7 | -5 | 1022 | SE | 6.25 | 2 | 0 |
| 30 | 2010 | 1 | 2 | 5 | 109 | -7 | -6 | 1022 | SE | 7.14 | 3 | 0 |

In [6]:

```
library(tidyverse)
library(lubridate)
```

```
Registered S3 method overwritten by 'rvest':
  method            from
  read_xml.response xml2
── Attaching packages ─────────────────────────────────────── tidyve
rse 1.2.1 ──
✔ tibble  2.1.1       ✔ dplyr   0.8.0.1
✔ readr   1.3.1       ✔ stringr 1.4.0
✔ purrr   0.3.3       ✔ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_co
nflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
✖ dplyr::select() masks MASS::select()

Attaching package: ‘lubridate’

The following object is masked from ‘package:base’:

    date
```

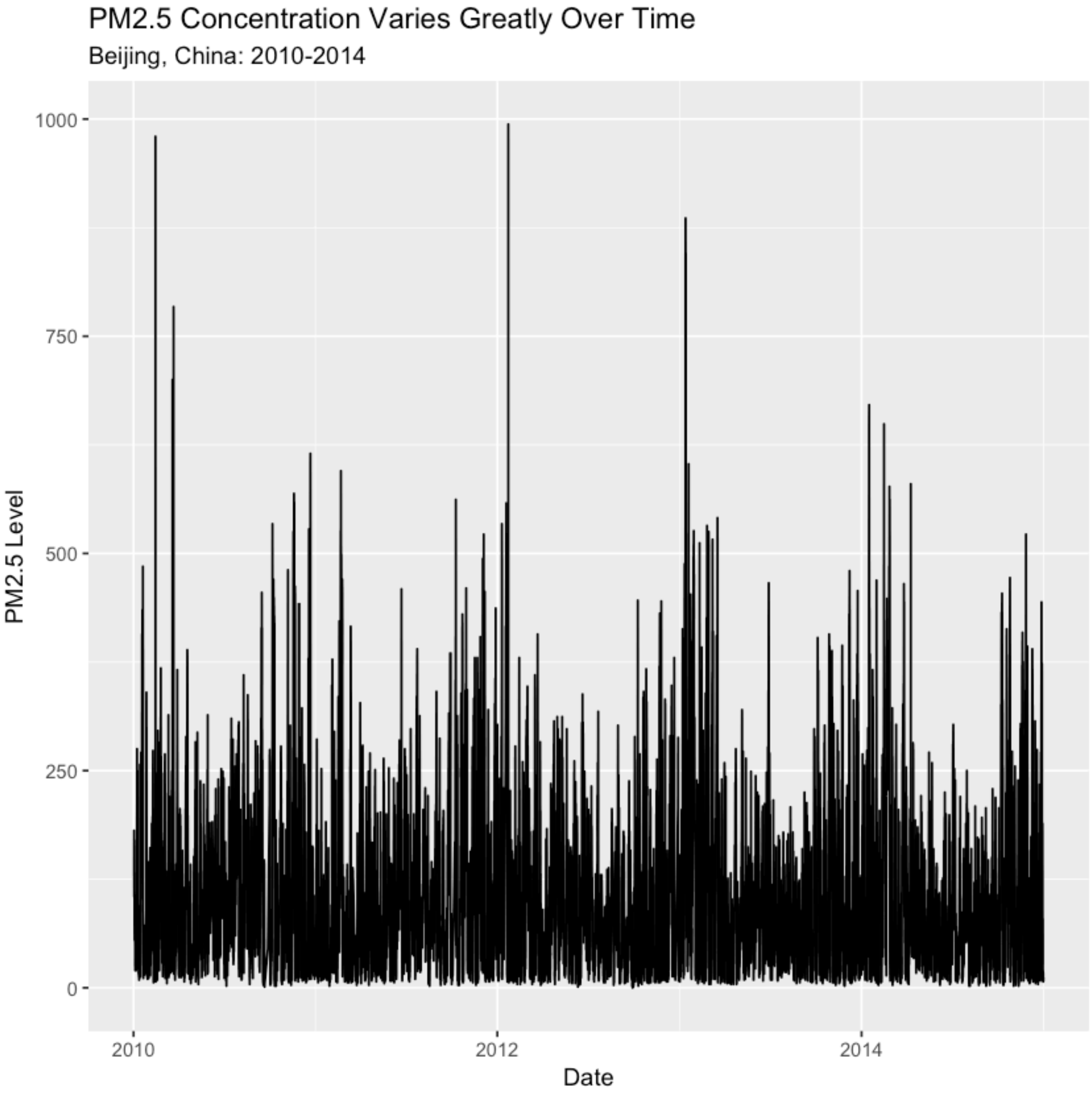
In [7]:

```
#editing the formate of the data for plotting
quality2 = quality2 %>%
  mutate(date = make_date(year, month, day))

quality2$datetime = as.POSIXct(paste(quality2$date,
        quality2$hour), format="%Y-%m-%d %H")

quality3 = quality2[-c(1,2,3,4,5,10, 14)]

head(quality3)
```

| pm2.5 | dewp | temp | pres | iws | is | ir | datetime |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 129 | -16 | -4 | 1020 | 1.79 | 0 | 0 | 2010-01-02 00:00:00 |
| 148 | -15 | -4 | 1020 | 2.68 | 0 | 0 | 2010-01-02 01:00:00 |
| 159 | -11 | -5 | 1021 | 3.57 | 0 | 0 | 2010-01-02 02:00:00 |
| 181 | -7 | -5 | 1022 | 5.36 | 1 | 0 | 2010-01-02 03:00:00 |
| 138 | -7 | -5 | 1022 | 6.25 | 2 | 0 | 2010-01-02 04:00:00 |
| 109 | -7 | -6 | 1022 | 7.14 | 3 | 0 | 2010-01-02 05:00:00 |

In [8]:

```
ggplot(data = quality3, aes(x = datetime, y = pm2.5)) +
  geom_line() +
  labs(x = "Date", y = "PM2.5 Level",
    title = "PM2.5 Concentration Varies Greatly Over Time",
    subtitle = "Beijing, China: 2010-2014")
```



In [9]:

```
#creating new dataframe for use in summarising statistics
quality4 = quality3[-c(8)]
head(quality4)
```

| pm2.5 | dewp | temp | pres | iws | is | ir |
|---|---|---|---|---|---|---|
| 129 | -16 | -4 | 1020 | 1.79 | 0 | 0 |
| 148 | -15 | -4 | 1020 | 2.68 | 0 | 0 |
| 159 | -11 | -5 | 1021 | 3.57 | 0 | 0 |
| 181 | -7 | -5 | 1022 | 5.36 | 1 | 0 |
| 138 | -7 | -5 | 1022 | 6.25 | 2 | 0 |
| 109 | -7 | -6 | 1022 | 7.14 | 3 | 0 |

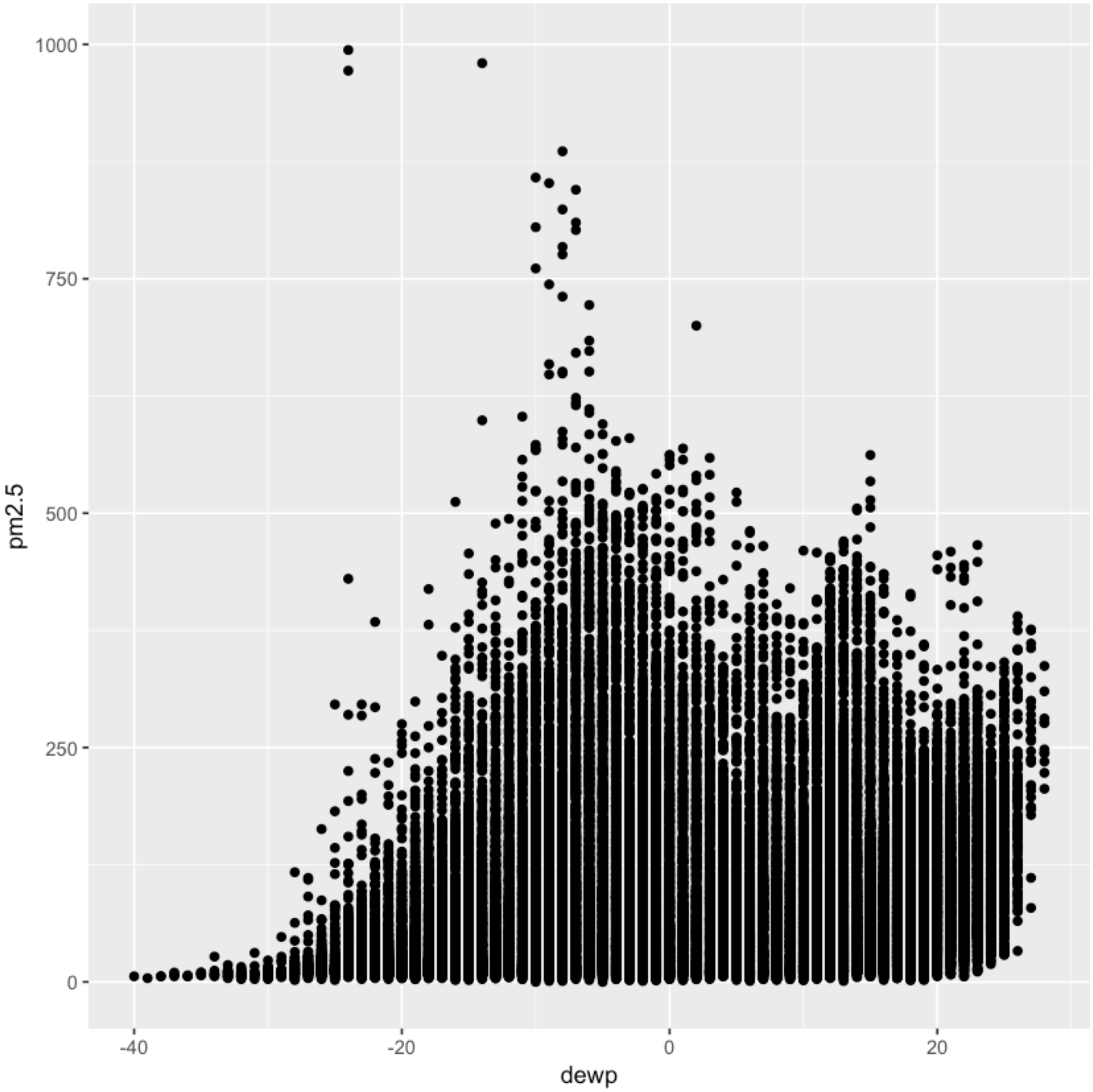
In [10]:

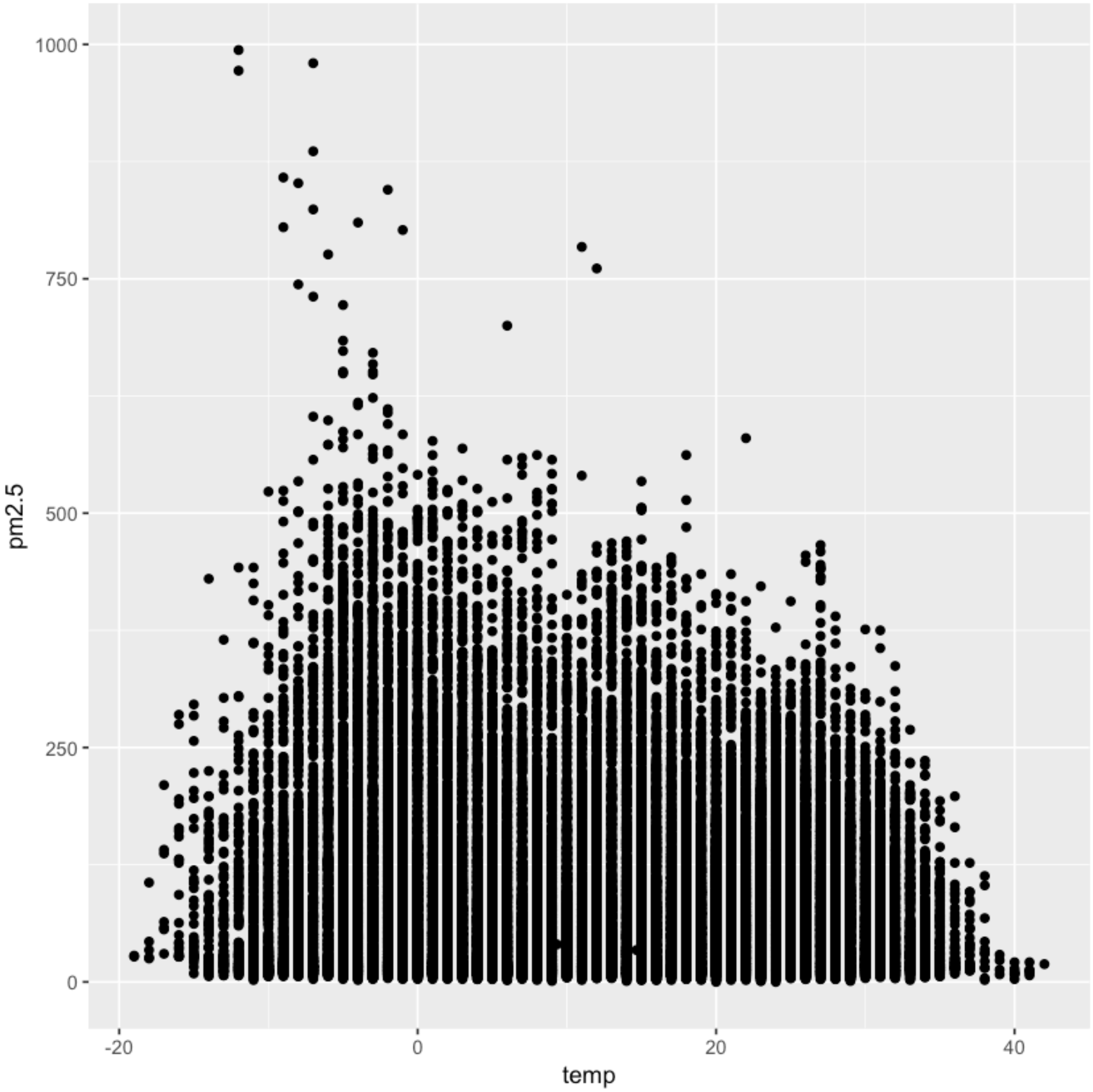
```
summary(quality4)
```

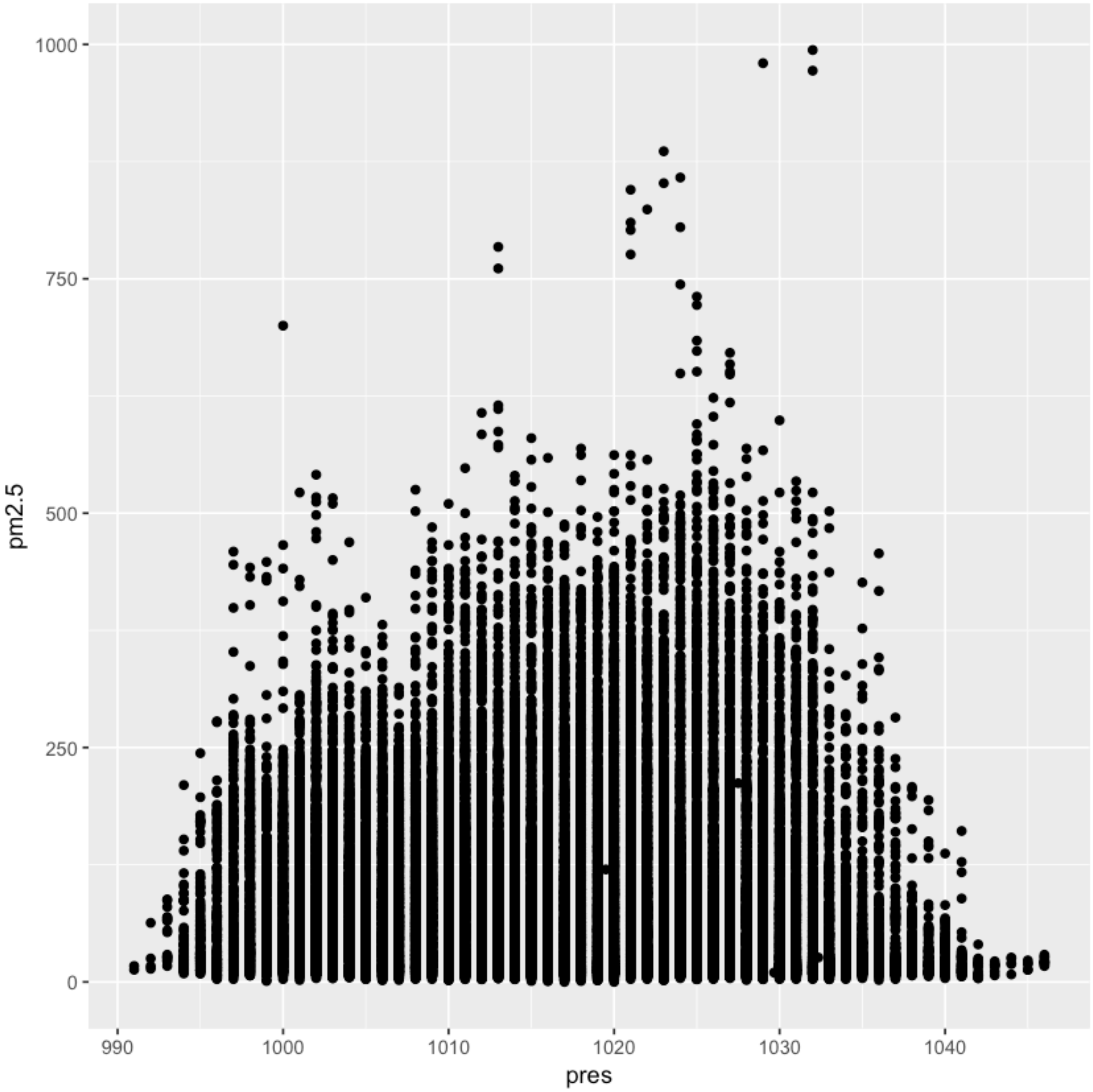
```
     pm2.5             dewp              temp             pres     
 Min.   :  0.00   Min.   :-40.00   Min.   :-19.0   Min.   : 991  
 1st Qu.: 29.00   1st Qu.:-10.00   1st Qu.:  2.0   1st Qu.:1008  
 Median : 72.00   Median :  2.00   Median : 14.0   Median :1016  
 Mean   : 98.61   Mean   :  1.75   Mean   : 12.4   Mean   :1016  
 3rd Qu.:137.00   3rd Qu.: 15.00   3rd Qu.: 23.0   3rd Qu.:1025  
 Max.   :994.00   Max.   : 28.00   Max.   : 42.0   Max.   :1046  
      iws               is                 ir         
 Min.   :  0.45   Min.   : 0.00000   Min.   : 0.0000  
 1st Qu.:  1.79   1st Qu.: 0.00000   1st Qu.: 0.0000  
 Median :  5.37   Median : 0.00000   Median : 0.0000  
 Mean   : 23.87   Mean   : 0.05534   Mean   : 0.1949  
 3rd Qu.: 21.91   3rd Qu.: 0.00000   3rd Qu.: 0.0000  
 Max.   :565.49   Max.   :27.00000   Max.   :36.0000  
```

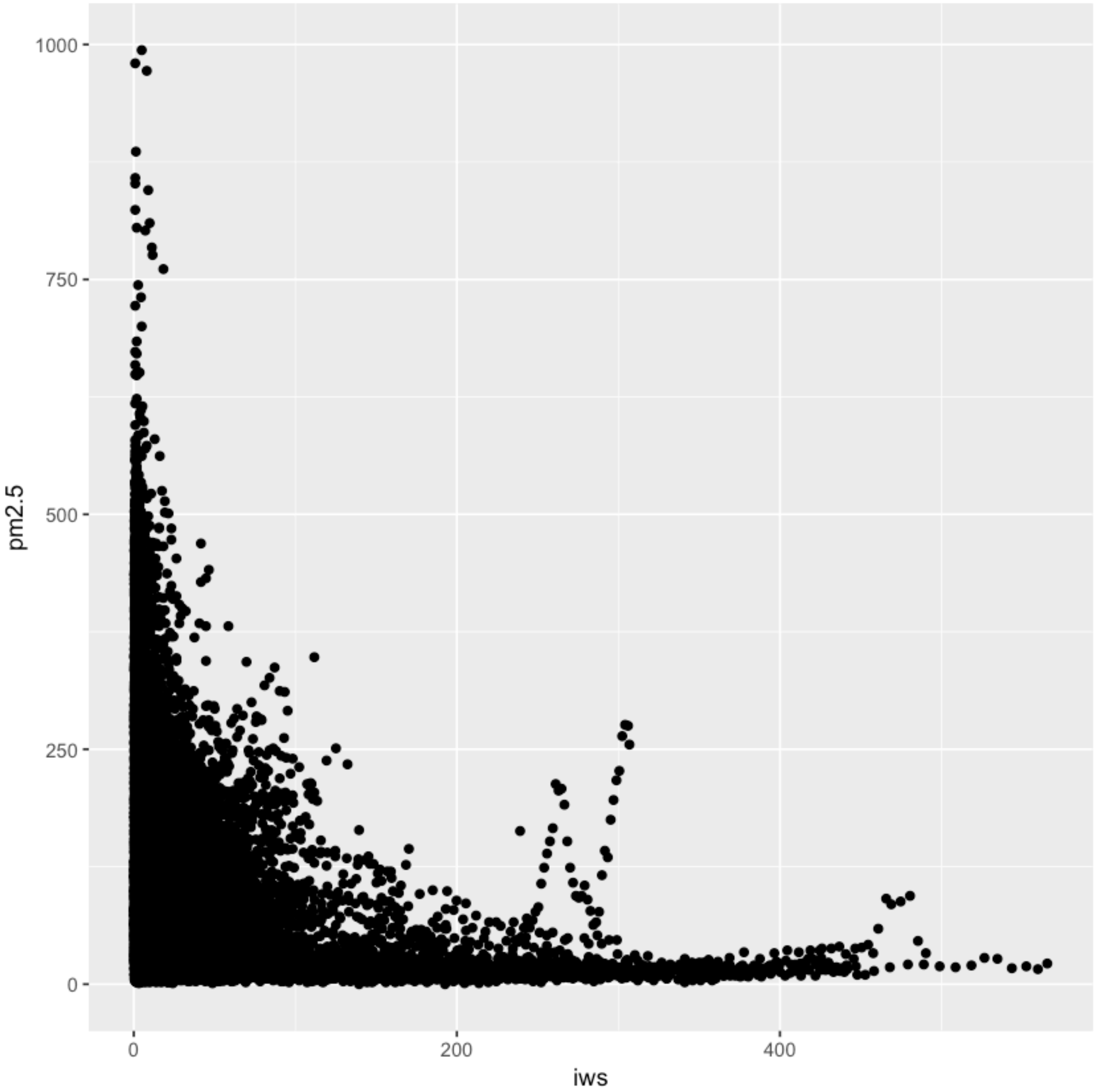
In [11]:

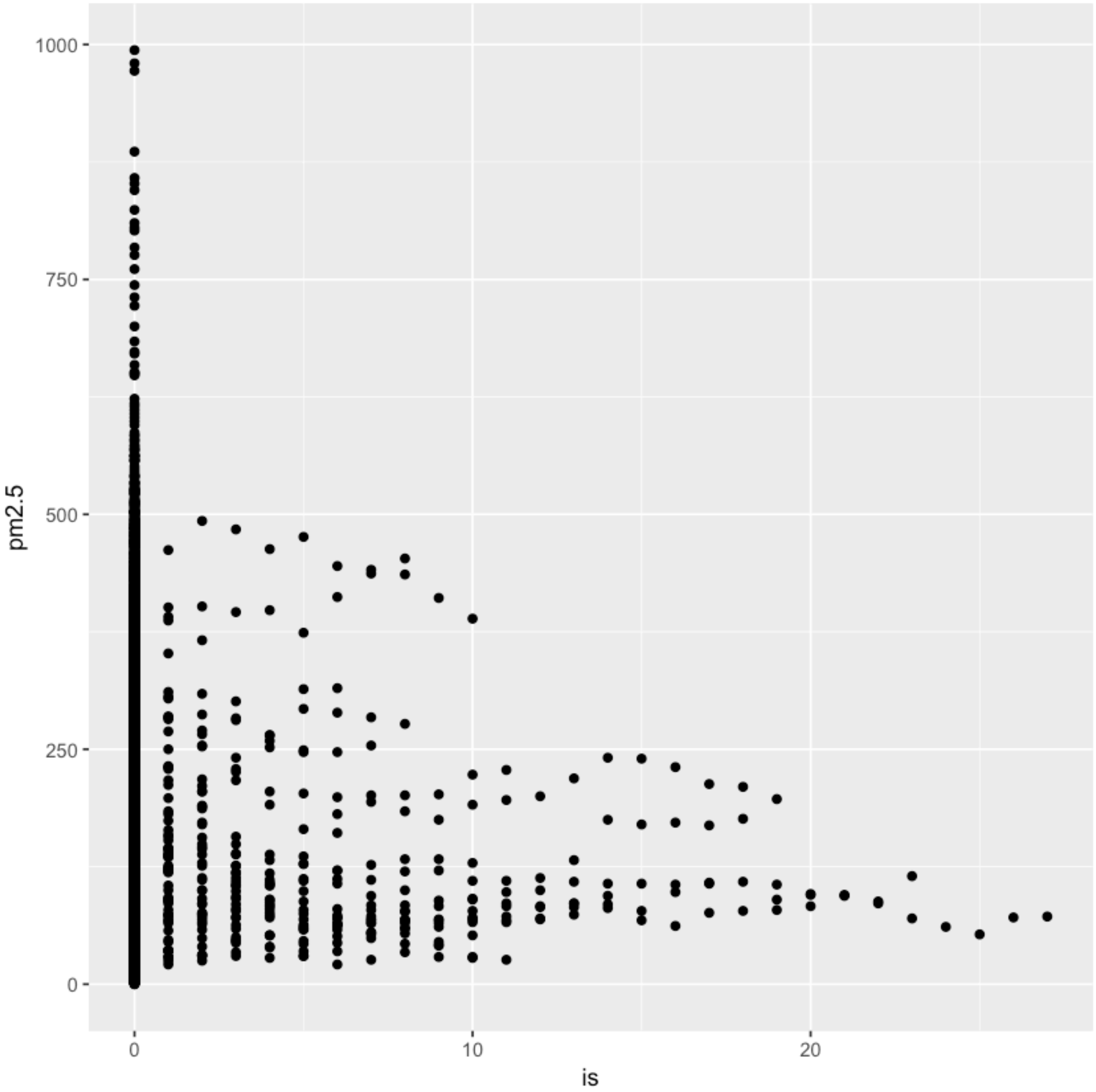
```
#plotting each variable vs. pm2.5 levels
par(mfrow = c(3,2), col.axis = "white", col.lab = "white", tck = 0)

ggplot(data = quality4, aes(x = dewp, y = pm2.5)) + geom_point()
ggplot(data = quality4, aes(x = temp, y = pm2.5)) + geom_point()
ggplot(data = quality4, aes(x = pres, y = pm2.5)) + geom_point()
ggplot(data = quality4, aes(x = iws, y = pm2.5)) + geom_point()
ggplot(data = quality4, aes(x = is, y = pm2.5)) + geom_point()
ggplot(data = quality4, aes(x = ir, y = pm2.5)) + geom_point()

#ggarrange(dewp, temp, pres, iws, is, ir, ncol = 3, nrow = 2)
```

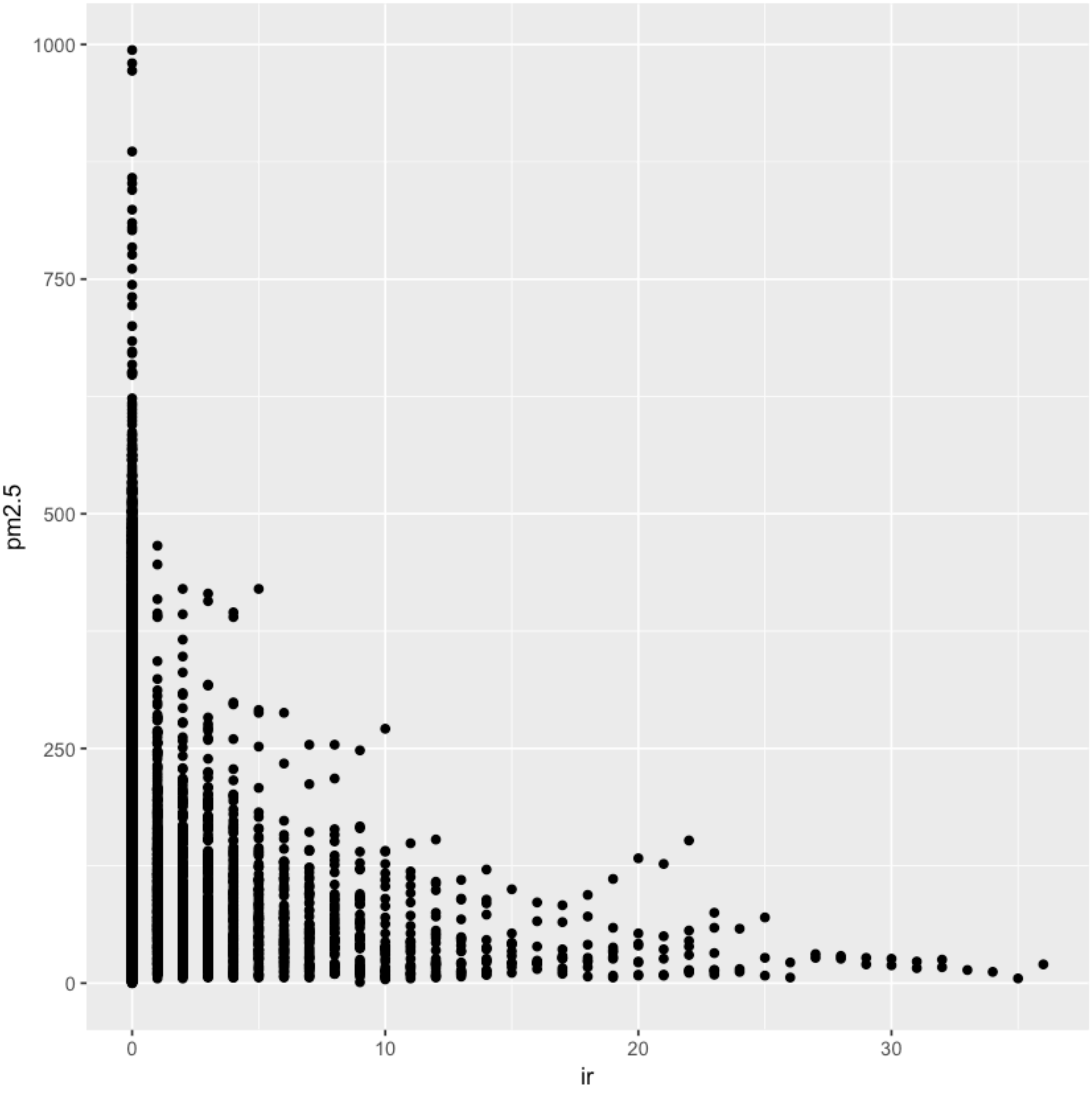

1000
750
500
250
0
pm2.5
-40
-20
0
20
dewp


1000
750
500
250
0
pm2.5
-20
0
20
40
temp


1000
750
500
250
0
pm2.5
990
1000
1010
1020
1030
1040
pres


1000
750
500
250
0
pm2.5
0
200
400
iws


1000
750
500
250
0
pm2.5
0
10
20
is



In [12]:

```
#creating initial model
quality.lm = lm(pm2.5 ~  dewp + temp +
                pres + iws + is + ir, data = quality4)
summary(quality.lm)
anova(quality.lm)
```

```
Call:
lm(formula = pm2.5 ~ dewp + temp + pres + iws + is + ir, data = qual
ity4)

Residuals:
    Min      1Q  Median      3Q     Max 
-163.00  -52.31  -15.68   33.09  874.89 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.728e+03  7.299e+01  23.680  < 2e-16 ***
dewp         4.282e+00  5.346e-02  80.109  < 2e-16 ***
temp        -6.068e+00  6.836e-02 -88.764  < 2e-16 ***
pres        -1.529e+00  7.135e-02 -21.431  < 2e-16 ***
iws         -2.616e-01  8.436e-03 -31.015  < 2e-16 ***
is          -2.267e+00  5.097e-01  -4.448  8.7e-06 ***
ir          -7.206e+00  2.816e-01 -25.593  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.46 on 41750 degrees of freedom
Multiple R-squared:  0.2361,	Adjusted R-squared:  0.236 
F-statistic:  2151 on 6 and 41750 DF,  p-value: < 2.2e-16
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **dewp** | 1 | 10397039.4 | 10397039.366 | 1606.12865 | 0.000000e+00 |
| **temp** | 1 | 59142103.7 | 59142103.736 | 9136.23809 | 0.000000e+00 |
| **pres** | 1 | 2868883.7 | 2868883.672 | 443.18350 | 7.097183e-98 |
| **iws** | 1 | 6796714.5 | 6796714.523 | 1049.95254 | 1.666189e-227 |
| **is** | 1 | 102830.7 | 102830.737 | 15.88523 | 6.741610e-05 |
| **ir** | 1 | 4239916.8 | 4239916.839 | 654.97991 | 2.363577e-143 |
| **Residuals** | 41750 | 270262531.2 | 6473.354 | NA | NA |

In [13]:

```
#calculating F-statistic for initial model
qf(.95, df1 = 6, df2 = 41750)
```

2.09881381597824

In [14]:

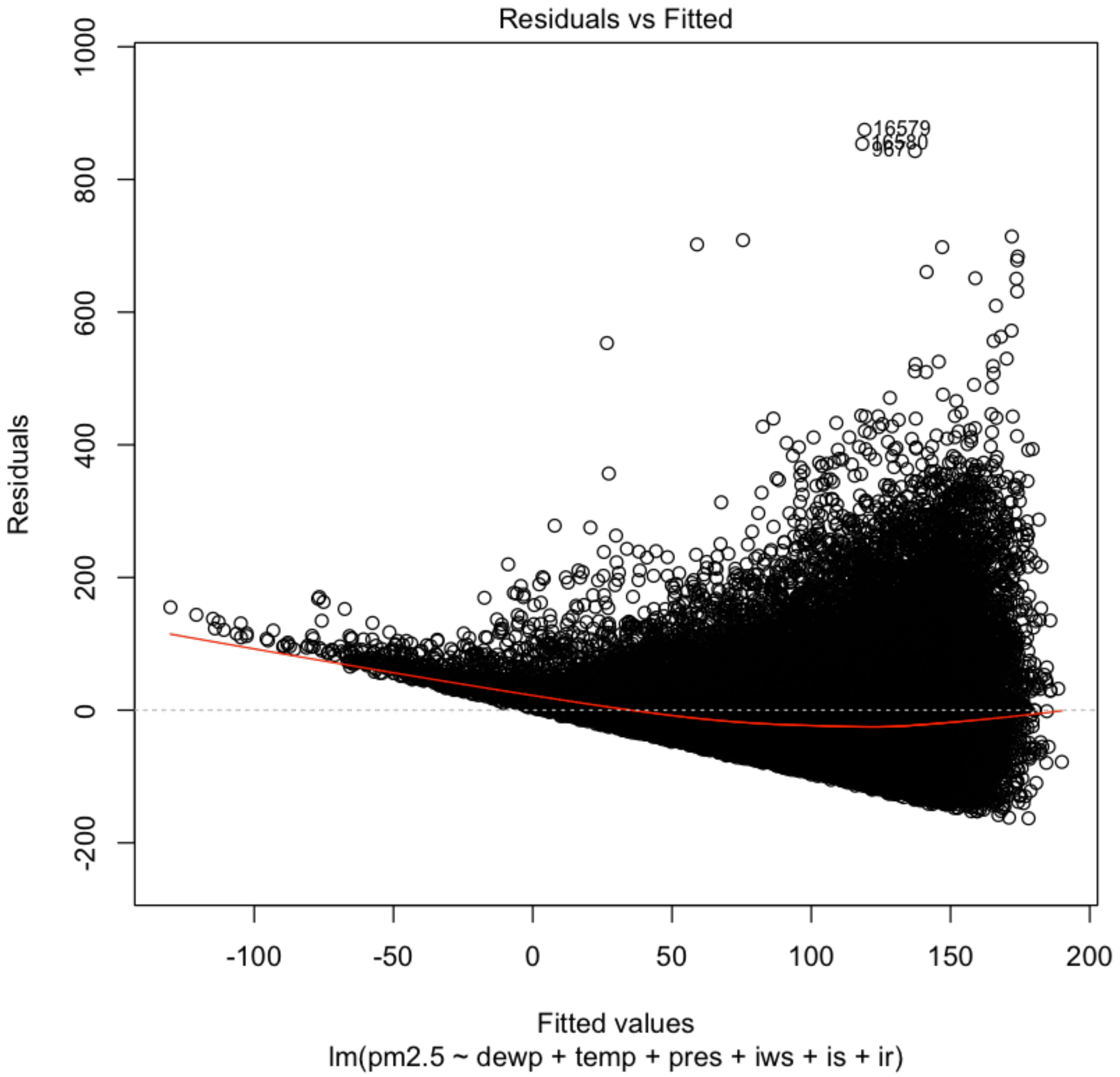
```
#creating a matrix in order to calculate t-values for each variable and critical t for the model
n = 41757
p = 7
quality4['B0'] = rep(1, 41757)
y = matrix(quality4$pm2.5, ncol = 1)
X = matrix(c(quality4$B0,quality4$dewp, quality4$temp, quality4$pres, quality4$iws,
            quality4$is, quality4$ir), ncol = 7, byrow = FALSE)
beta.hat = solve(t(X)%*%X)%*%t(X)%*%y
SSres = as.vector(t(y)%*%y - t(beta.hat)%*%t(X)%*%y)
sig.hat = SSres/(n-p)
SSreg = as.vector(t(beta.hat)%*%t(X)%*%y - n*mean(y)^2)
C = solve(t(X)%*%X)
beta.se = sqrt(sig.hat*diag(C))
t = beta.hat/beta.se
critical.t = qt(1-0.5/2, n-p)
abs(t)
critical.t
```

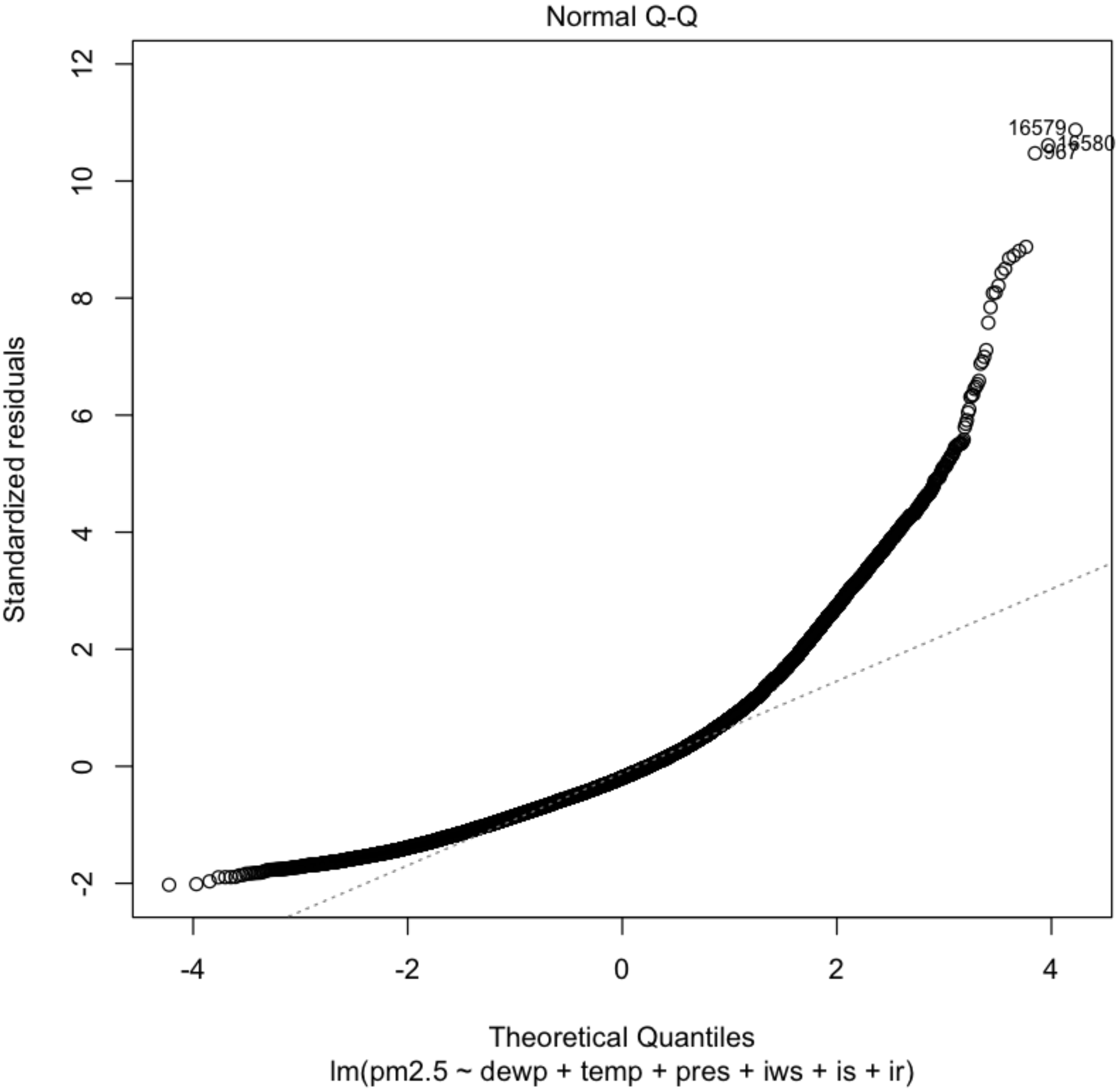
```
23.680393

80.108929

88.764250

21.431247

31.014553

 4.447625

25.592575
```

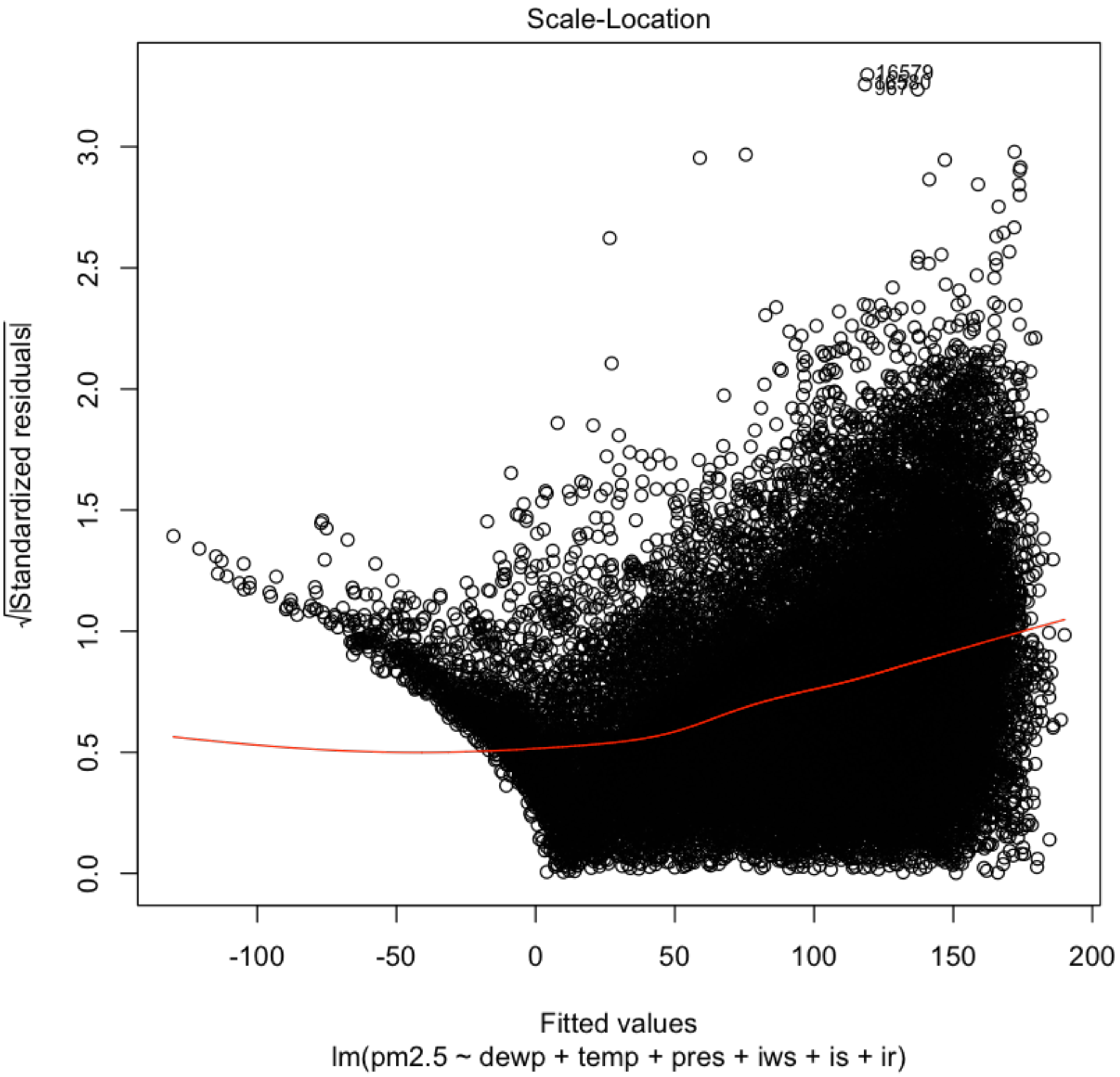
```
0.674495626527342
```

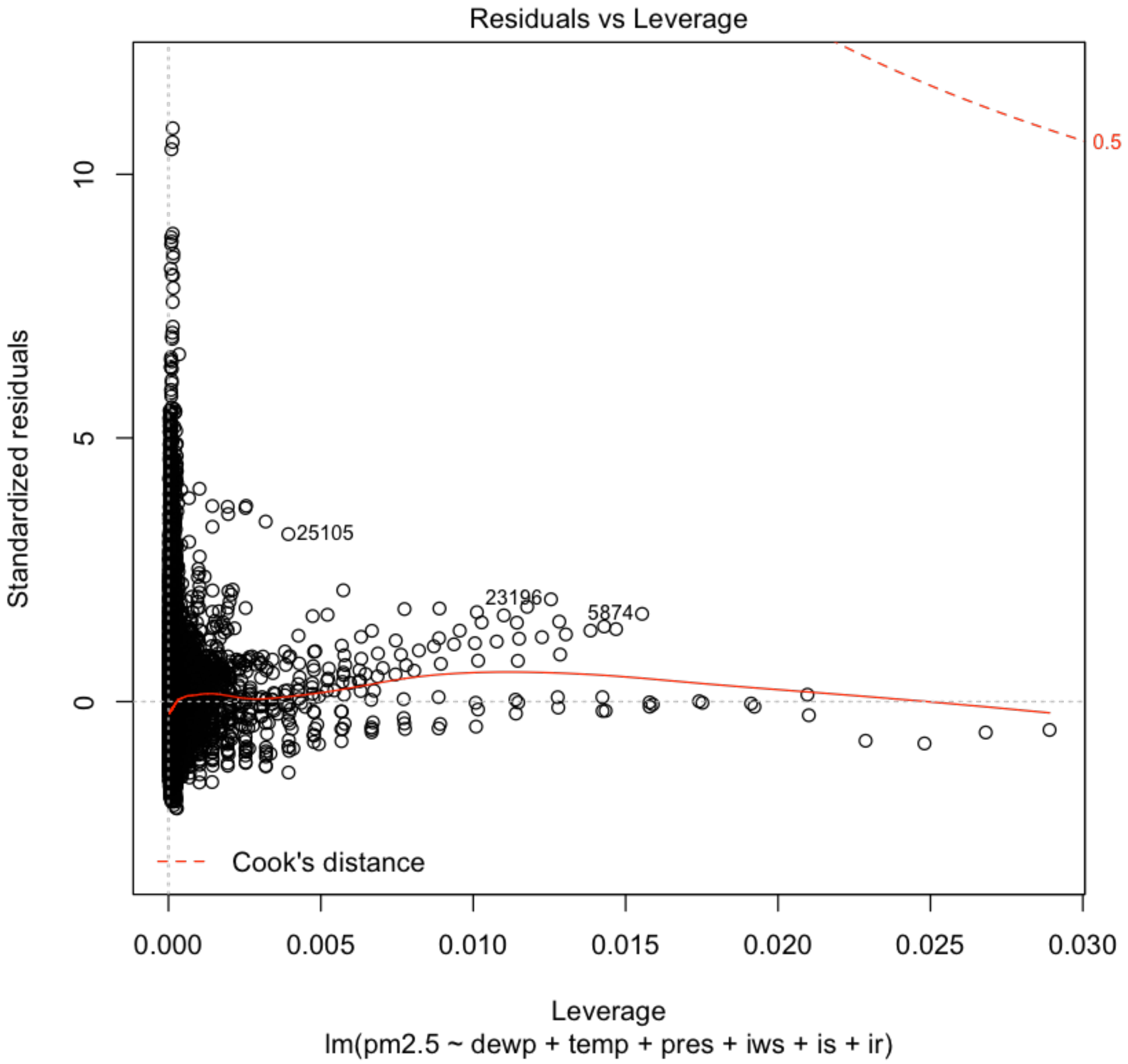
In [45]:

```
#plotting normal and residual vs fitted plots for the full model
plot(quality.lm)
```


Residuals vs Fitted
16579
16580
967
Residuals
1000
800
600
400
200
0
-200
-100
-50
0
50
100
150
200
Fitted values
lm(pm2.5 ~ dewp + temp + pres + iws + is + ir)


Normal Q-Q
Standardized residuals
12
10
8
6
4
2
0
-2
16579
16580
967
-4
-2
0
2
4
Theoretical Quantiles
lm(pm2.5 ~ dewp + temp + pres + iws + is + ir)

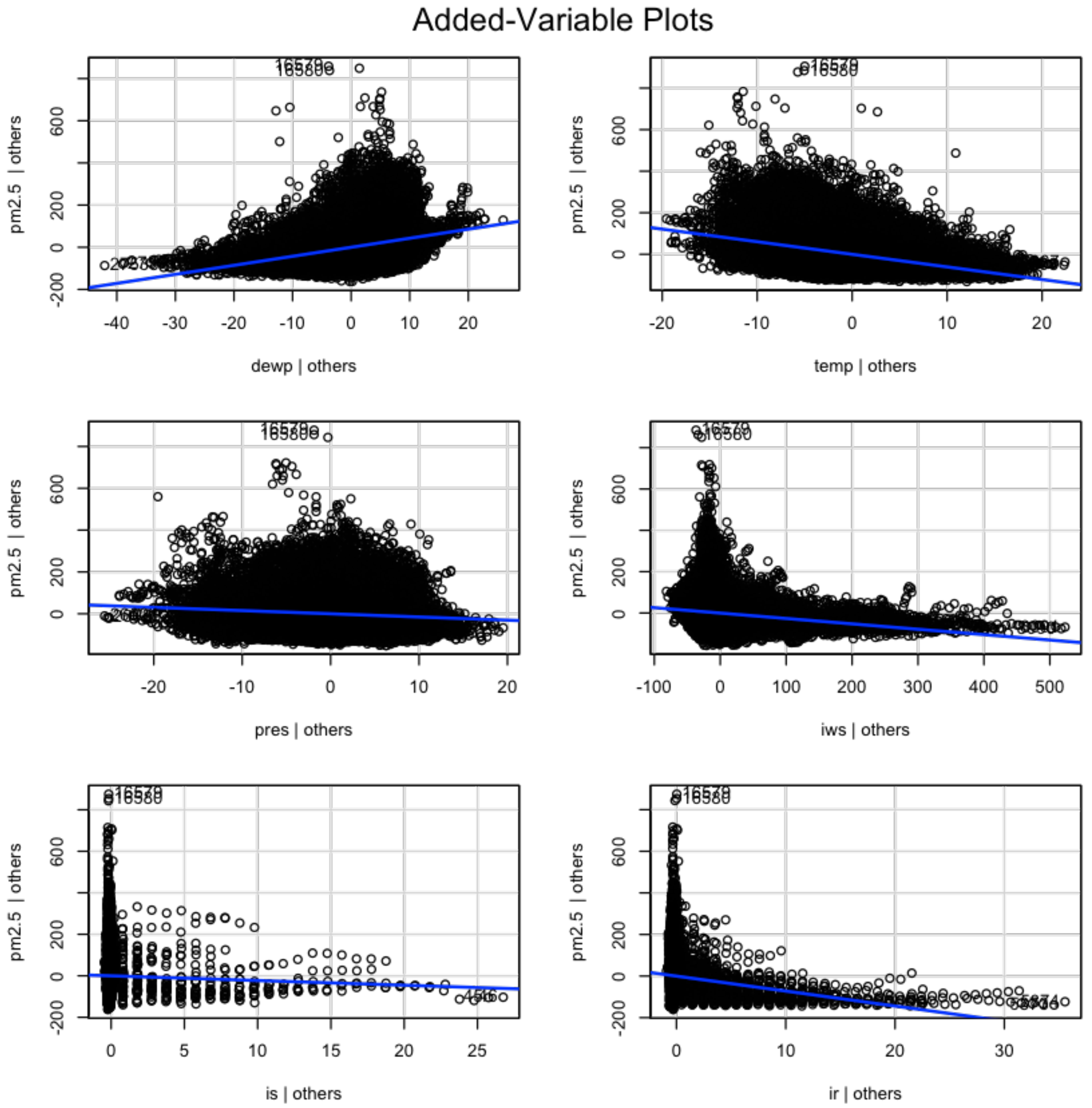

Scale-Location
16579
16580
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-100
-50
0
50
100
150
200
Fitted values
lm(pm2.5 ~ dewp + temp + pres + iws + is + ir)



Residuals vs Leverage
Standardized residuals
10
5
0
0.5
25105
23196
5874
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(pm2.5 ~ dewp + temp + pres + iws + is + ir)

In [130]:

```
avPlots(quality.lm)
```



In [46]:

```
#looking at residuals relationships
studentized.residuals = rstandard(quality.lm)
r.student.residuals = rstudent(quality.lm)
residuals = data.frame(studentized.residuals, r.student.residuals)
residuals$ID = seq.int(nrow(quality2))
head(residuals)
```

| studentized.residuals | r.student.residuals | ID |
|---:|---:|---:|
| 0.06219617 | 0.06219543 | 1 |
| 0.24802944 | 0.24802665 | 2 |
| 0.11832494 | 0.11832354 | 3 |
| 0.23187989 | 0.23187727 | 4 |
| -0.27154686 | -0.27154385 | 5 |
| -0.67645439 | -0.67645000 | 6 |

In [48]:

```
#claculating min in order to transform the model appropriately based on log transformation
min(quality2$temp)
min(quality2$dewp)
min(quality2$pm2.5)
min(quality2$ir)
min(quality2$is)
min(quality2$iws)
min(quality2$pres)
```

```
-19

-40

0

0

0

0.45

991
```

In [52]:

```
# creating logarithmic transformed model
quality.lm2 = lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1)
+ log(iws) + log(is + 1) + log(pres), data = quality2)#log(temp + 20) + log(dewp
+ 41), data = quality2)
summary(quality.lm2)
```

```
Call:
lm(formula = log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) +
    log(ir + 1) + log(iws) + log(is + 1) + log(pres), data = quality
2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9139 -0.5299  0.0624  0.5947  4.6837

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    66.246346   4.878310  13.580   <2e-16 ***
log(temp + 20) -1.364319   0.017341 -78.677   <2e-16 ***
log(dewp + 41)  1.863805   0.019710  94.561   <2e-16 ***
log(ir + 1)    -0.317148   0.012597 -25.176   <2e-16 ***
log(iws)       -0.137644   0.002805 -49.069   <2e-16 ***
log(is + 1)     0.024739   0.023314   1.061    0.289
log(pres)      -9.250335   0.697051 -13.271   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8196 on 41750 degrees of freedom
Multiple R-squared:  0.3358,    Adjusted R-squared:  0.3357
F-statistic:  3518 on 6 and 41750 DF,  p-value: < 2.2e-16
```

In [53]:

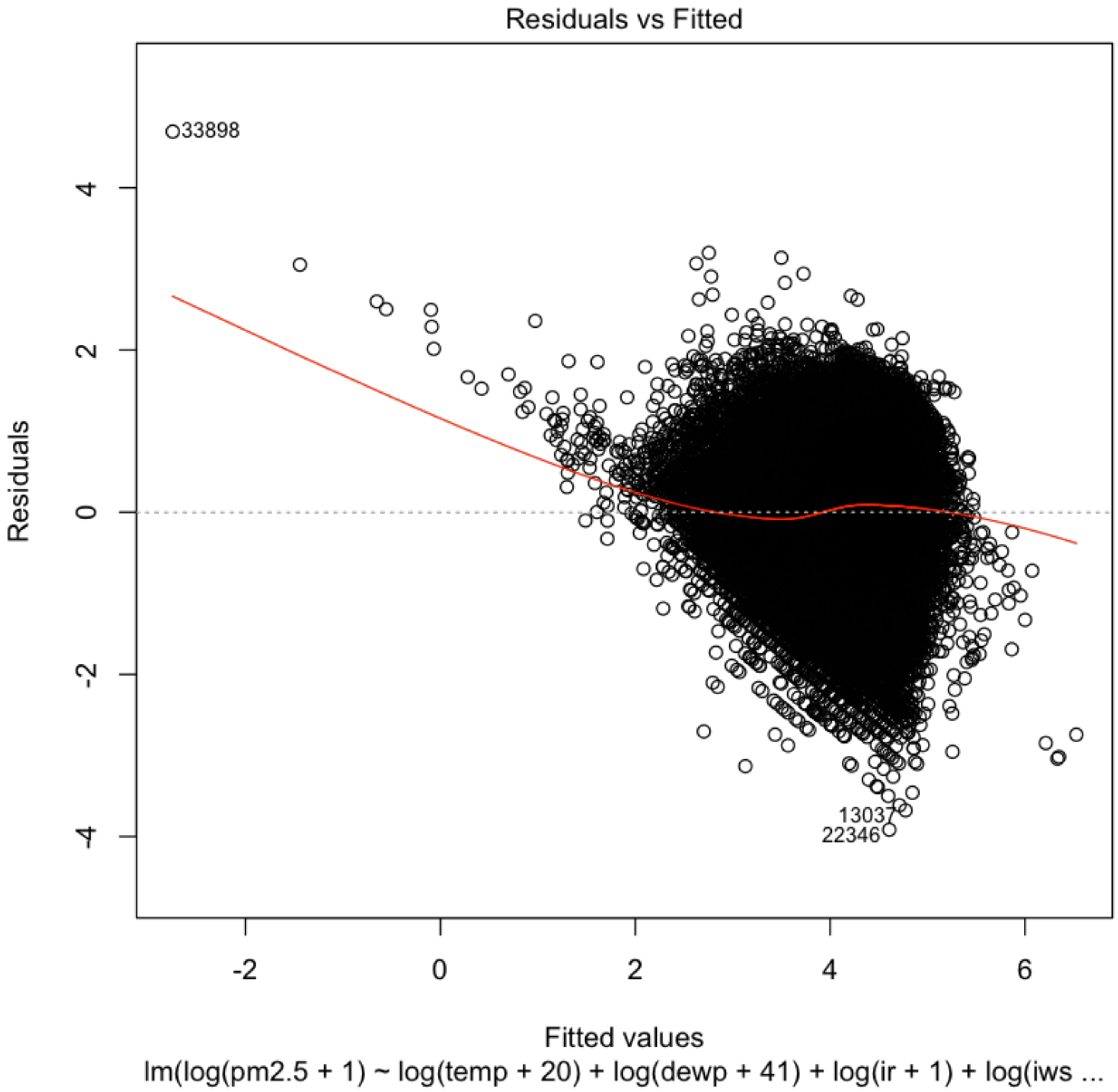
```
# creating log transformed and partial model
quality.lm3 = lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1)
+
        log(iws) + log(pres),
        data = quality2)#log(temp + 20) + log(dewp + 41), data = quality2)
summary(quality.lm3)
anova(quality.lm3)
```

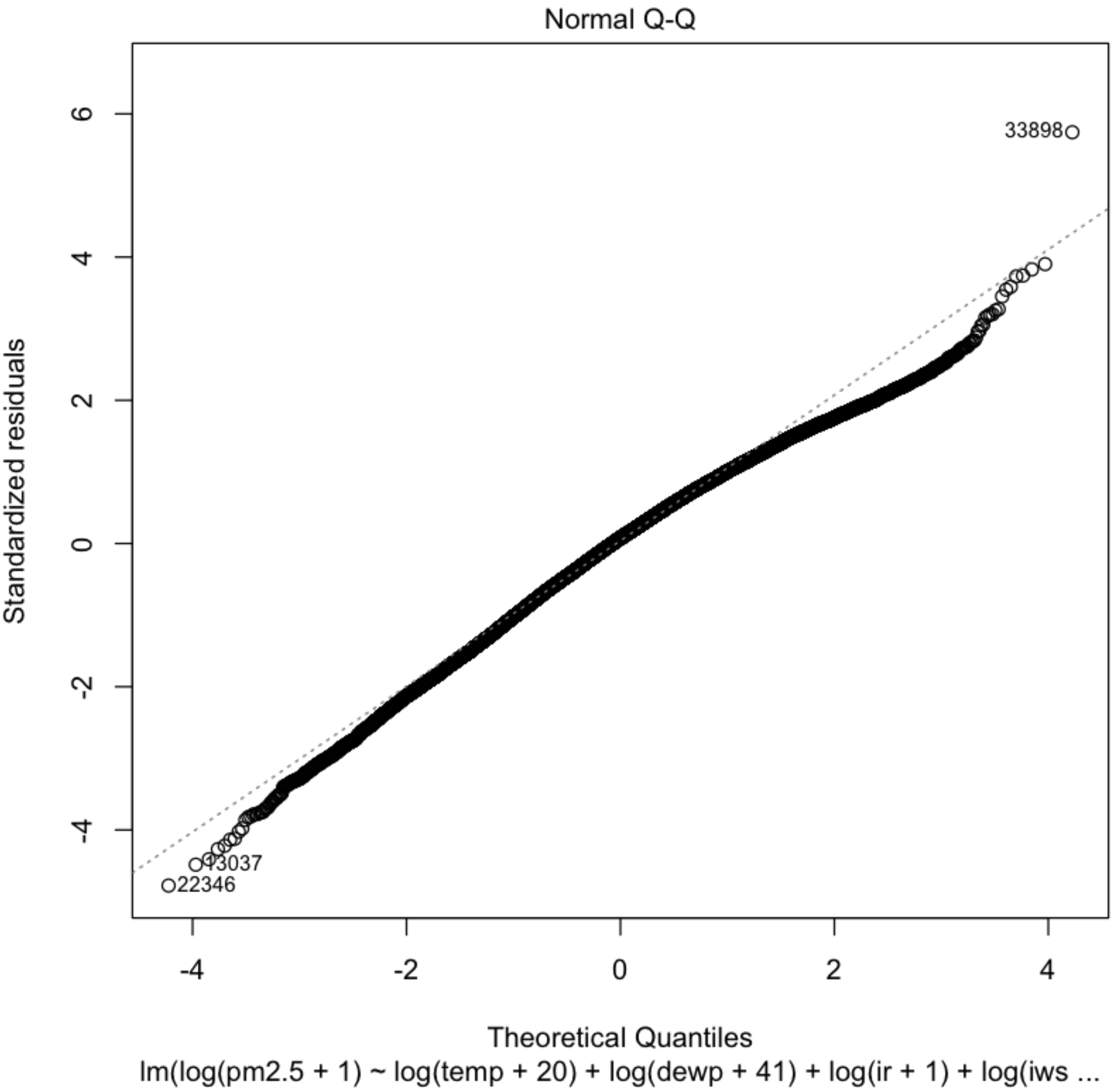
```
Call:
lm(formula = log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + 
    log(ir + 1) + log(iws) + log(pres), data = quality2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9147 -0.5297  0.0626  0.5943  4.6923 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    66.046737   4.874689   13.55   <2e-16 ***
log(temp + 20) -1.366840   0.017177  -79.57   <2e-16 ***
log(dewp + 41)  1.866705   0.019520   95.63   <2e-16 ***
log(ir + 1)    -0.317493   0.012593  -25.21   <2e-16 ***
log(iws)       -0.137392   0.002795  -49.16   <2e-16 ***
log(pres)      -9.221829   0.696534  -13.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8196 on 41751 degrees of freedom
Multiple R-squared:  0.3358,    Adjusted R-squared:  0.3357 
F-statistic:  4221 on 5 and 41751 DF,  p-value: < 2.2e-16
```

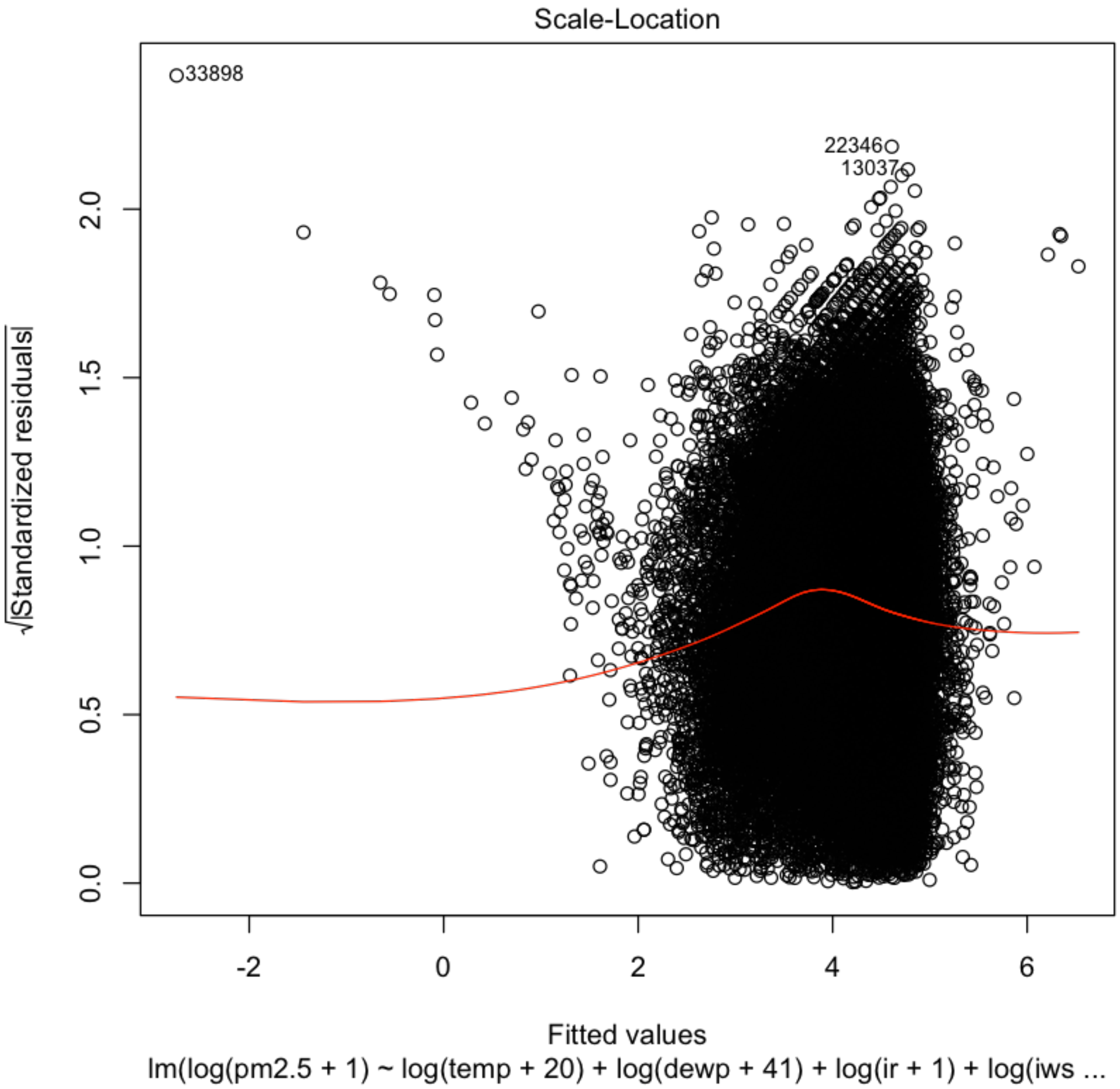
| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---:|---:|---:|---:|---:|---:|
| **log(temp + 20)** | 1 | 21.48726 | 2.148726e+01 | 31.9896 | 1.560116e-08 |
| **log(dewp + 41)** | 1 | 11928.36177 | 1.192836e+04 | 17758.5943 | 0.000000e+00 |
| **log(ir + 1)** | 1 | 512.46088 | 5.124609e+02 | 762.9367 | 1.949623e-166 |
| **log(iws)** | 1 | 1597.19997 | 1.597200e+03 | 2377.8644 | 0.000000e+00 |
| **log(pres)** | 1 | 117.73934 | 1.177393e+02 | 175.2869 | 6.239461e-40 |
| **Residuals** | 41751 | 28043.94445 | 6.716952e-01 | NA | NA |

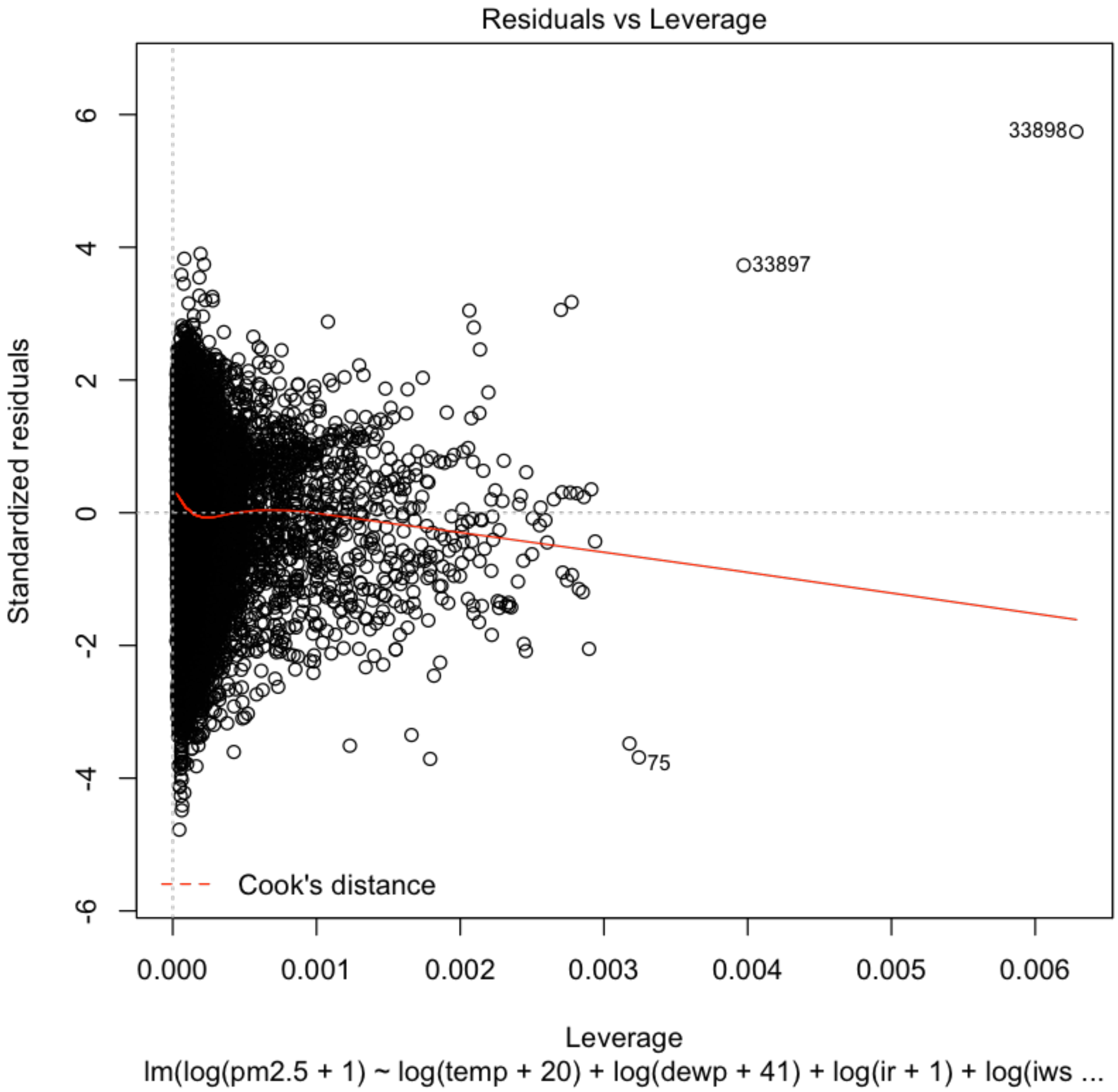
In [57]:

```
#plotting residual vs fitted and normal plot for the partial, transformed model
plot(quality.lm3)
```


Residuals vs Fitted
Residuals
4
2
0
-2
-4
33898
13037
22346
-2
0
2
4
6
Fitted values
lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1) + log(iws ...

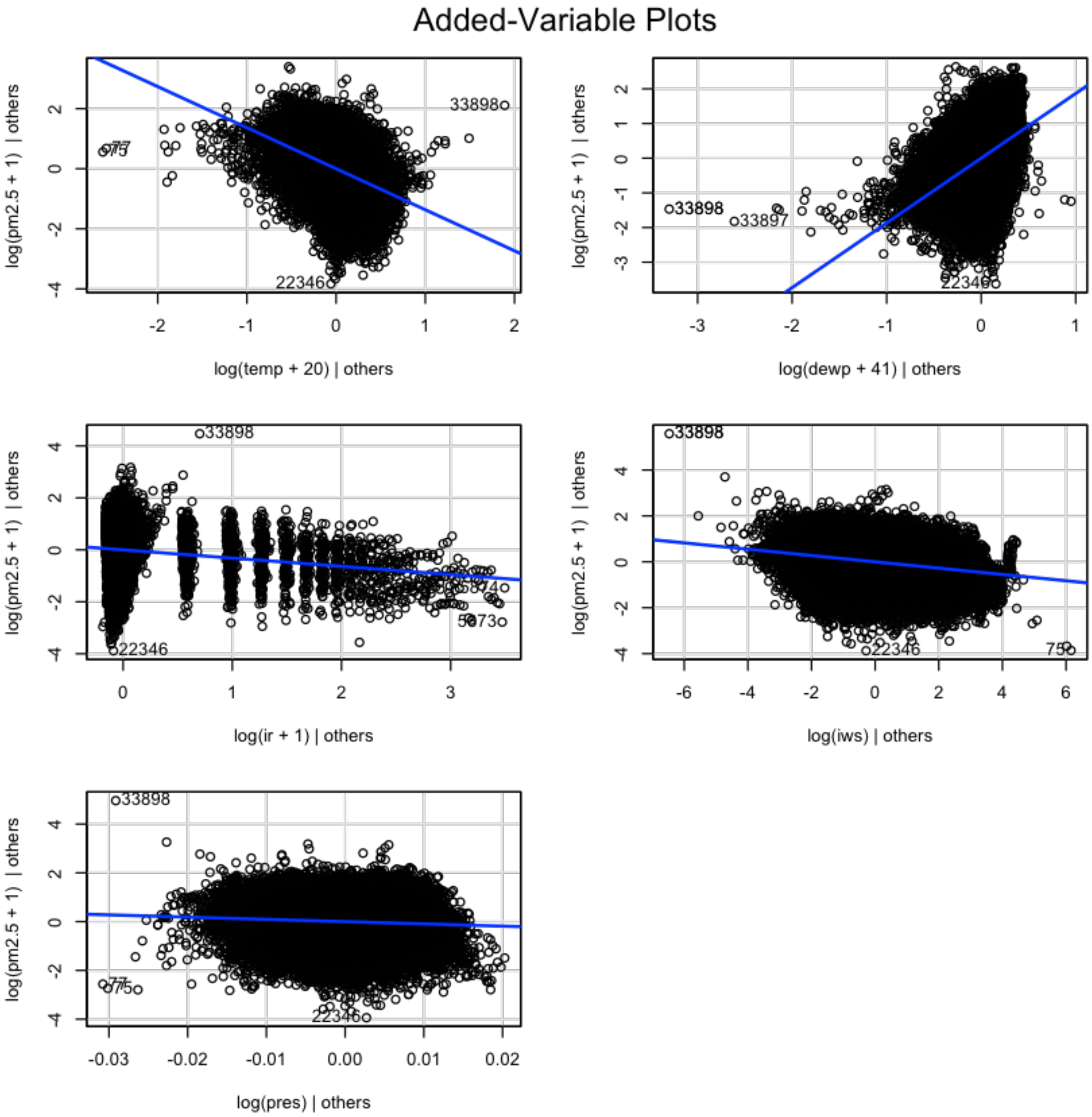

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1) + log(iws ...
6
4
2
0
-2
-4
-4
-2
0
2
4
33898
13037
22346


Scale-Location
33898
22346
13037
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
-2
0
2
4
6
Fitted values
lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1) + log(iws ...


Residuals vs Leverage
Standardized residuals
6
4
2
0
-2
-4
-6
0.000
0.001
0.002
0.003
0.004
0.005
0.006
Leverage
lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1) + log(iws ...
33898
33897
75
Cook's distance

In [60]:

```
avPlots(quality.lm3)
```



In [68]:

```
#calculating PRESS statistics for each model as an indicator of predictive accuracy
full = sum((resid(quality.lm) / (1-lm.influence(quality.lm)$hat))^2)
full
trans1 = sum((resid(quality.lm2) / (1-lm.influence(quality.lm2)$hat))^2)
trans1
trans2 = sum((resid(quality.lm3) / (1-lm.influence(quality.lm3)$hat))^2)
trans2
```

```
270338852.262718

28052.5727840162

28052.5011265684
```

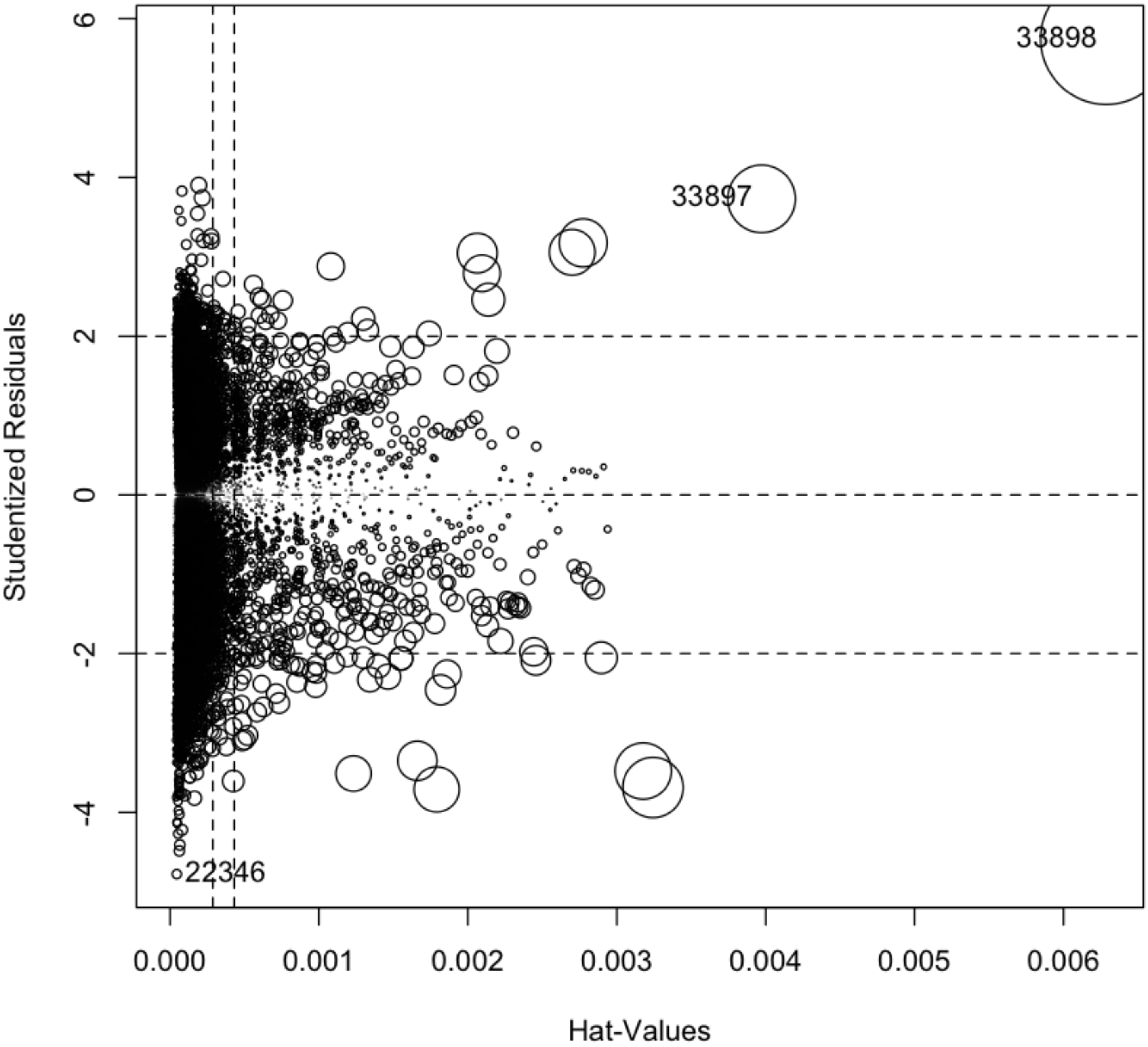
In [69]:

```
library(car)
```

In [70]:

```
# looking at influential points in the more accurate partial, transformed model
influencePlot(quality.lm3)
```

```
         StudRes          Hat         CookD
22346  -4.777892 4.571148e-05  0.0001738359
33897   3.730552 3.972530e-03  0.0092481860
33898   5.745626 6.286264e-03  0.0347793890
```


33898
33897
22346
Studentized Residuals
Hat-Values
6
4
2
0
-2
-4
0.000
0.001
0.002
0.003
0.004
0.005
0.006

In [91]:

```
#creating a dataframe to simplify elimination of influential points
quality2$cooksd = cooks.distance(quality.lm3)
quality2$outlier = ifelse(quality2$cooksd < 4/nrow(quality2),
                     'keep', 'delete')


nrow(quality2)

quality3 = quality2[!(quality2$outlier == 'delete'),]
nrow(quality3)

nrow(quality2) - nrow(quality3)
```

```
41757

40063

1694
```

In [78]:

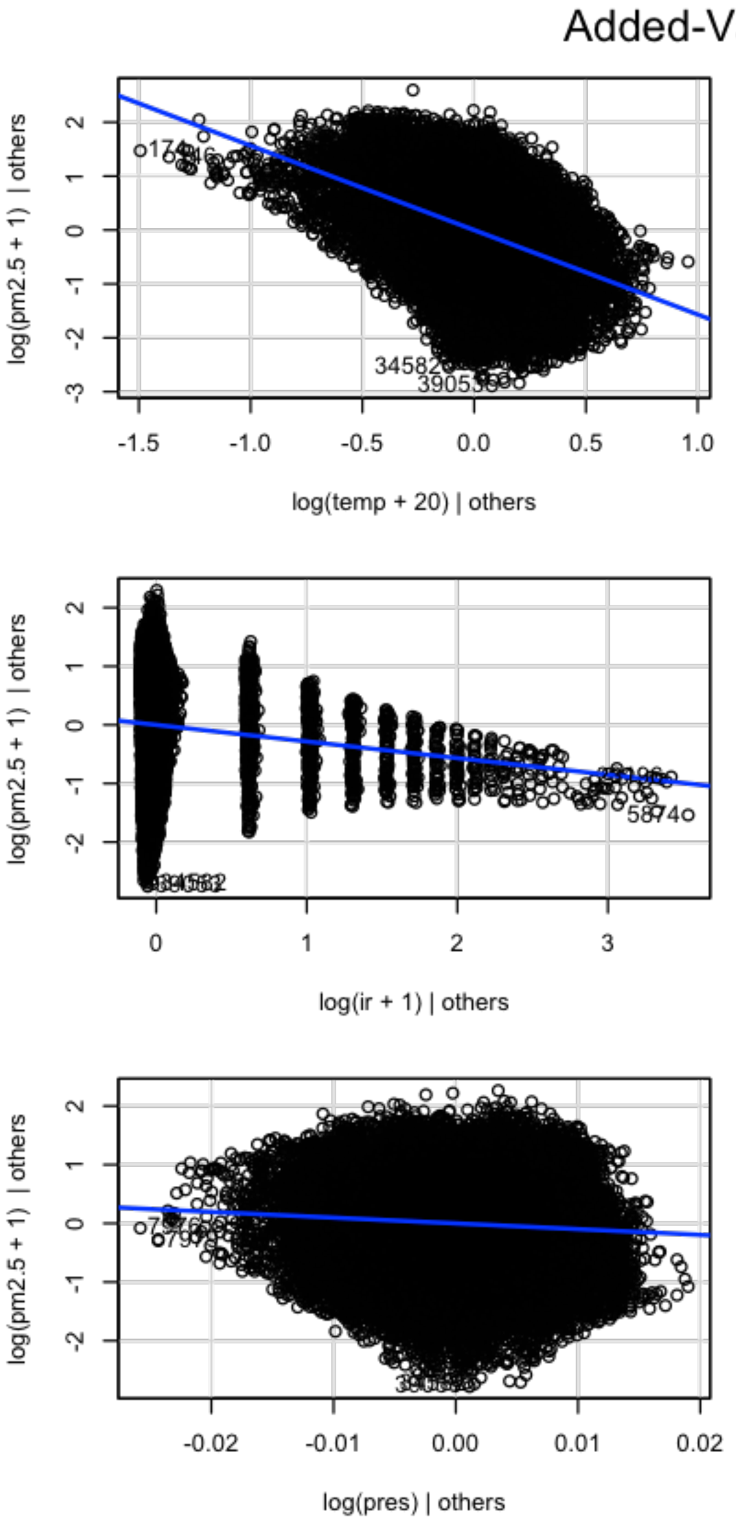
```
#creating new model, partial and transformed, using dataframe that has eliminate
d influential points
quality.lm4 = lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1)
+
        log(iws) + log(pres),
        data = quality3)#log(temp + 20) + log(dewp + 41), data = quality2)
summary(quality.lm4)
anova(quality.lm4)
```

```
Call:
lm(formula = log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + 
    log(ir + 1) + log(iws) + log(pres), data = quality3)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.77128 -0.50820  0.05001  0.55719  2.29667 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    69.915960   4.717281   14.82   <2e-16 ***
log(temp + 20) -1.567131   0.017143  -91.42   <2e-16 ***
log(dewp + 41)  2.061474   0.019344  106.57   <2e-16 ***
log(ir + 1)    -0.283592   0.015382  -18.44   <2e-16 ***
log(iws)       -0.133139   0.002675  -49.77   <2e-16 ***
log(pres)      -9.783992   0.673881  -14.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7584 on 40057 degrees of freedom
Multiple R-squared:  0.3887,	Adjusted R-squared:  0.3886 
F-statistic:  5094 on 5 and 40057 DF,  p-value: < 2.2e-16
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---:|---:|---:|---:|---:|---:|
| **log(temp + 20)** | 1 | 4.304299 | 4.304299e+00 | 7.483632 | 6.228975e-03 |
| **log(dewp + 41)** | 1 | 12897.795002 | 1.289780e+04 | 22424.641208 | 0.000000e+00 |
| **log(ir + 1)** | 1 | 228.072824 | 2.280728e+02 | 396.536869 | 8.324667e-88 |
| **log(iws)** | 1 | 1399.218430 | 1.399218e+03 | 2432.739182 | 0.000000e+00 |
| **log(pres)** | 1 | 121.242749 | 1.212427e+02 | 210.797672 | 1.216175e-47 |
| **Residuals** | 40057 | 23039.252650 | 5.751617e-01 | NA | NA |

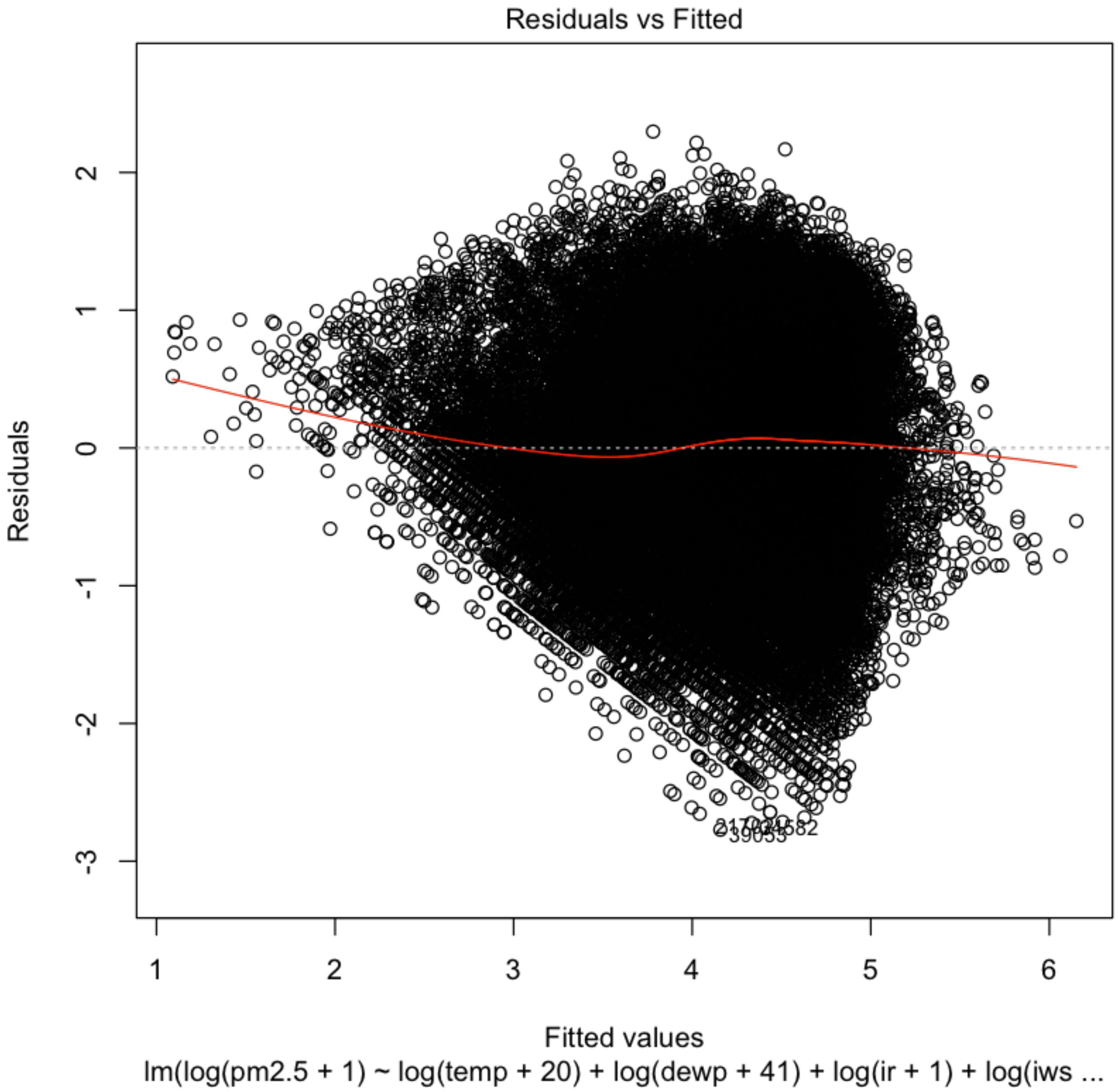
In [80]:

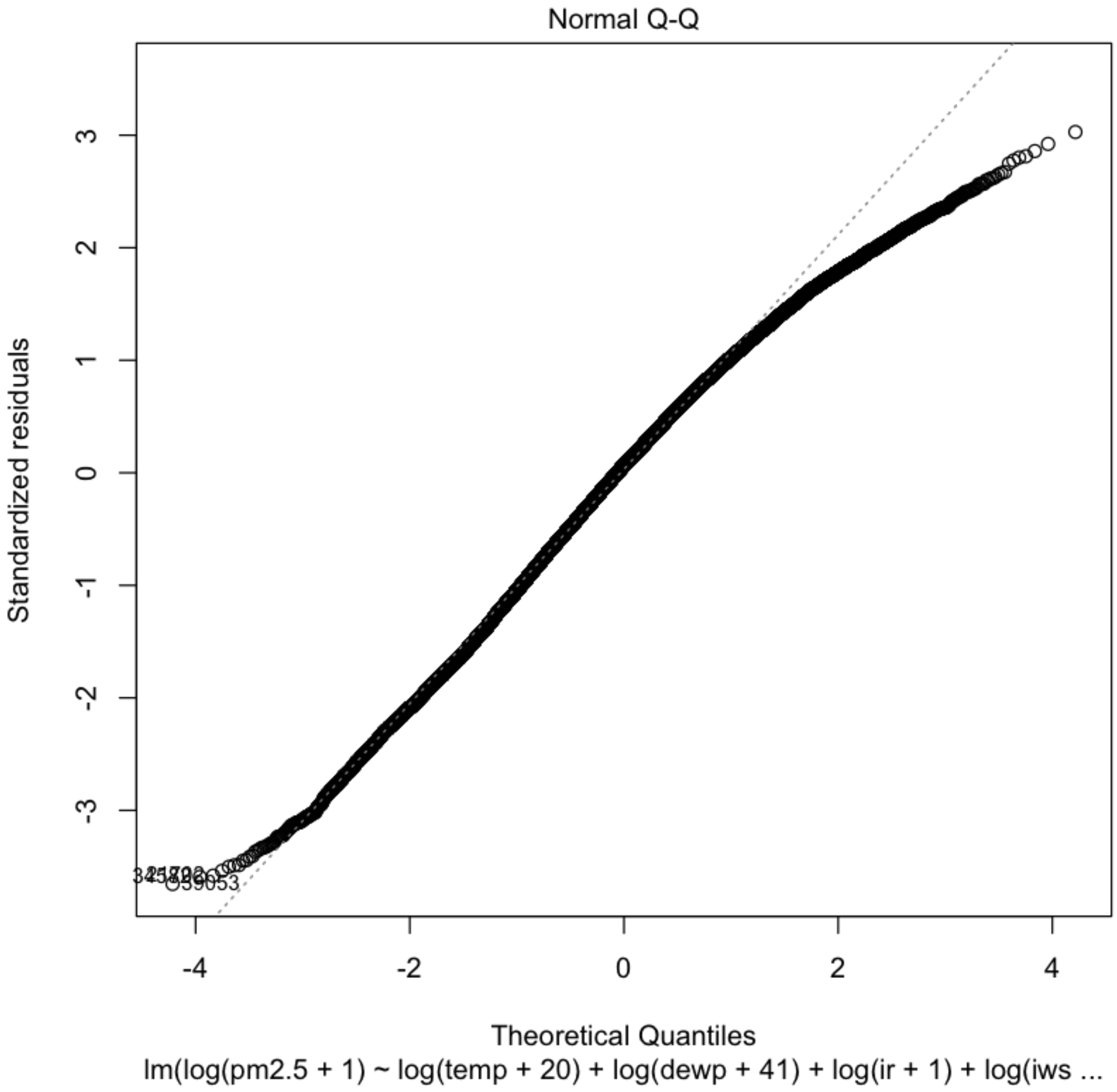
```
#viewing added variable plots for the final, partial model
avPlots(quality.lm4)
```

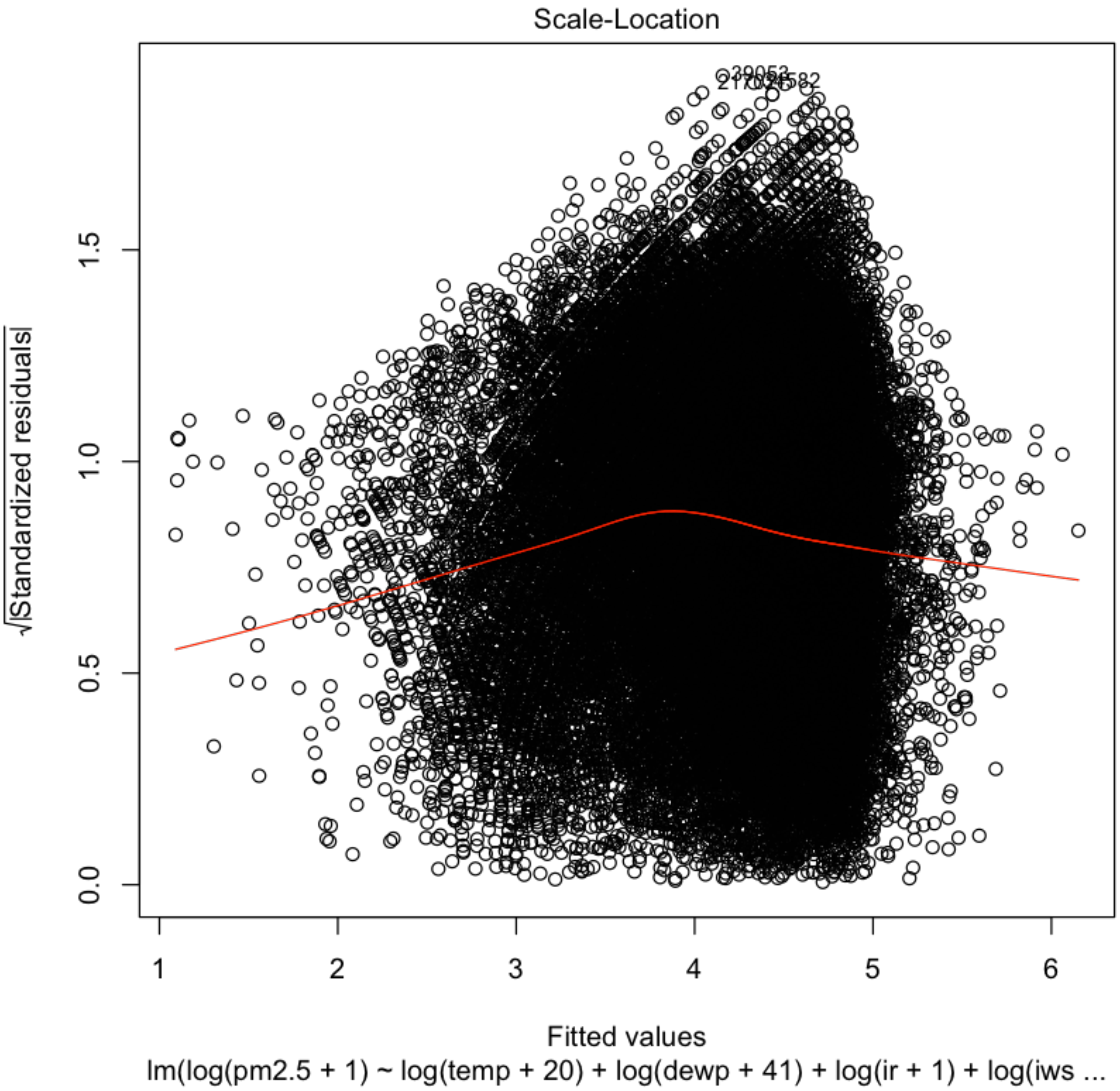


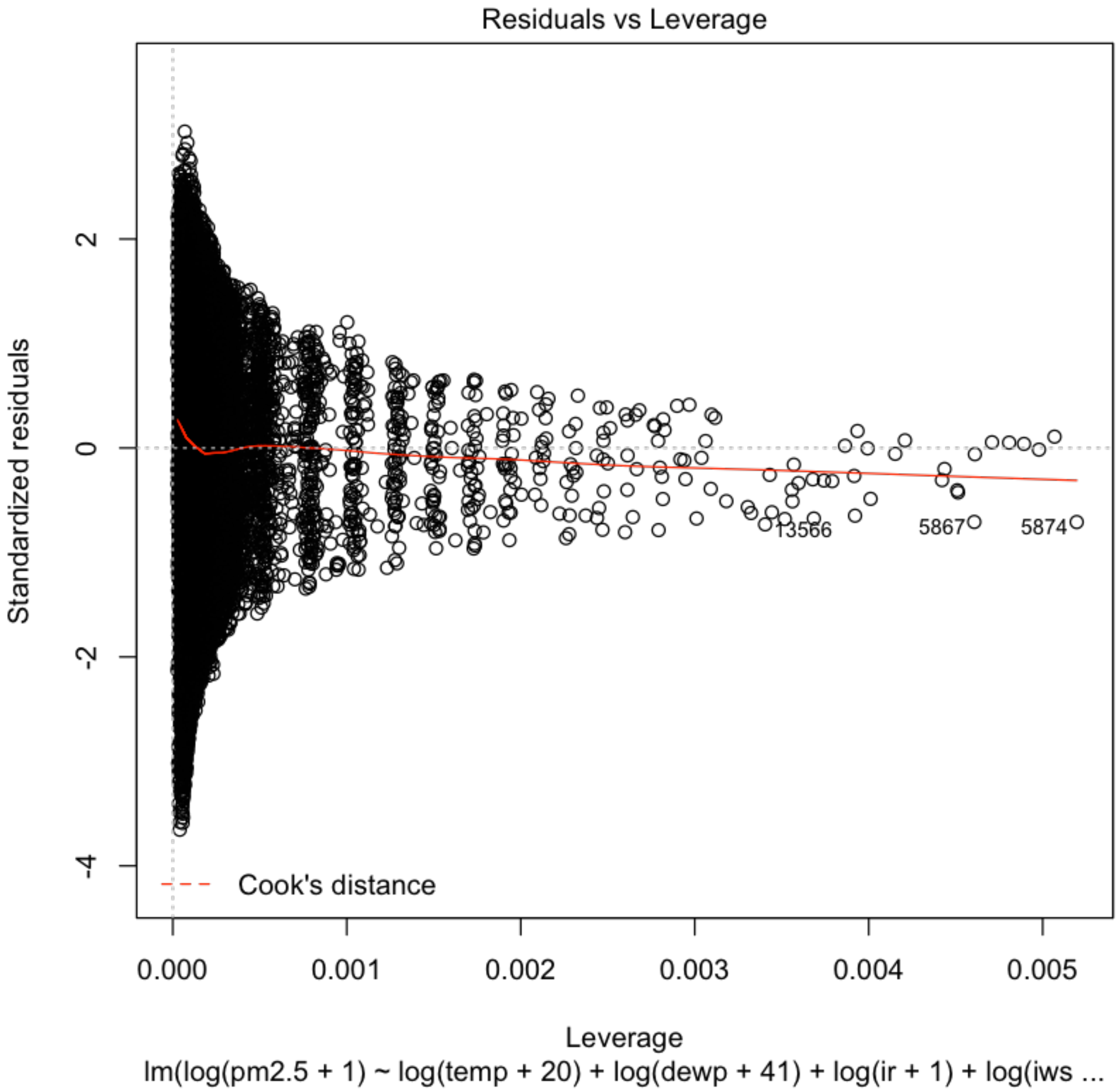
In [81]:

```
#plotting residual vs fitted and normal plot for the final model
plot(quality.lm4)
```


Residuals vs Fitted
Residuals
2
1
0
-1
-2
-3
217024582
39053
1
2
3
4
5
6
Fitted values
lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1) + log(iws ...


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
39053
-4
-2
0
2
4
Theoretical Quantiles
lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1) + log(iws ...

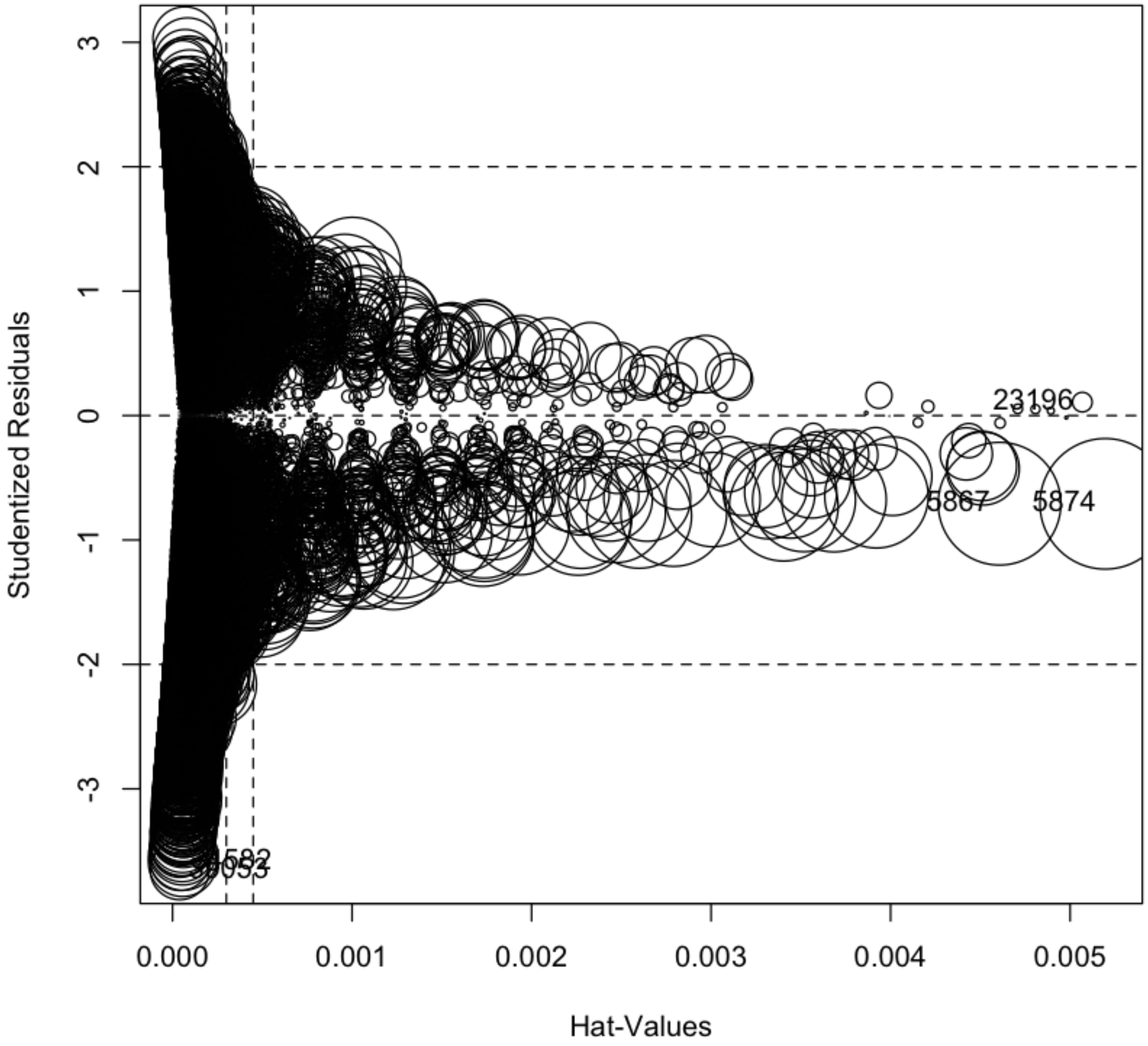

Scale-Location
29053
21702
4582
√|Standardized residuals|
0.0
0.5
1.0
1.5
1
2
3
4
5
6
Fitted values
lm(log(pm2.5 + 1) ~ log(temp + 20) + log(dewp + 41) + log(ir + 1) + log(iws ...



In [82]:

```
#calculating influential points and plot, to take a look at what has changed
influencePlot(quality.lm4)
```

|  | StudRes | Hat | CookD |
|---|---|---|---|
| **5867** | -0.7083800 | 4.606775e-03 | 3.870696e-04 |
| **5874** | -0.7093382 | 5.195548e-03 | 4.379802e-04 |
| **23196** | 0.1079260 | 5.068072e-03 | 9.889201e-06 |
| **34582** | -3.5905763 | 5.603128e-05 | 1.203658e-04 |
| **39053** | -3.6547811 | 3.993937e-05 | 8.889066e-05 |



In [84]:

```
#calculating PRESS statistic for the final model
trans.outliers = sum((resid(quality.lm4) / (1-lm.influence(quality.lm4)$hat))^2)
trans.outliers
```

23044.9043608328

In [86]:

```
#calculating F - statistic for the final model
qf(.95, df1 = 5, df2 = 40057)
```

```
2.21432259342585
```

In [93]:

```
quality3 = quality3[-c(1,2,3,4,5,10,12,14,15,16,17)]
head(quality3)
```

| pm2.5 | dewp | temp | pres | iws | ir |
|---:|---:|---:|---:|---:|---:|
| 129 | -16 | -4 | 1020 | 1.79 | 0 |
| 148 | -15 | -4 | 1020 | 2.68 | 0 |
| 159 | -11 | -5 | 1021 | 3.57 | 0 |
| 181 | -7 | -5 | 1022 | 5.36 | 0 |
| 138 | -7 | -5 | 1022 | 6.25 | 0 |
| 109 | -7 | -6 | 1022 | 7.14 | 0 |

In [95]:

```
#creating matrix from data to calculate t-values and critical t to determine sig
nificance.
n = 40063
p = 6
quality3['B0'] = rep(1, n)
y = matrix(quality3$pm2.5, ncol = 1)
X = matrix(c(quality3$B0,quality3$dewp, quality3$temp, quality3$pres, quality3$i
ws,
            quality3$ir), ncol = p, byrow = FALSE)
beta.hat = solve(t(X)%*%X)%*%t(X)%*%y
SSres = as.vector(t(y)%*%y - t(beta.hat)%*%t(X)%*%y)
sig.hat = SSres/(n-p)
SSreg = as.vector(t(beta.hat)%*%t(X)%*%y - n*mean(y)^2)
C = solve(t(X)%*%X)
beta.se = sqrt(sig.hat*diag(C))
t = beta.hat/beta.se
critical.t = qt(1-0.5/2, n-p)
abs(t)
critical.t
```

```
22.46548

83.59505

91.54862

20.17455

32.10869

21.34565
```

```
0.674495874891155
```

In [100]:

```
#calculting confidence interval on Betas in the final model
(confint(quality.lm4))
```

| | 2.5 % | 97.5 % |
|---|---|---|
| **(Intercept)** | 60.6699795 | 79.1619409 |
| **log(temp + 20)** | -1.6007315 | -1.5335308 |
| **log(dewp + 41)** | 2.0235587 | 2.0993897 |
| **log(ir + 1)** | -0.3137412 | -0.2534431 |
| **log(iws)** | -0.1383821 | -0.1278958 |
| **log(pres)** | -11.1048140 | -8.4631694 |