Short Report 2: Mortality and Pollution

Jeanny Zhang, Nina Sun

October 24, 2021

Introduction

This report is dedicated to investigating the effects of pollution on mortality accounting for the weather and demographic factors given the data collected in sixty cities in the US. According to the data set, the pollution was measured by NOX, SO2 and HC, and the weather and socioeconomic variables include precipitation, January temperature, percentage of 1960 population that is nonwhite, median number of school years completed by persons of age 25 years or more, and etc. A fitted multiple regression model is found in this report after model transformations, removing insignificant terms, and checking outliers, collinearity and assumptions. The report also discusses which of the pollution variable has the strongest association with mortality after accounting for weather and socioeconomic factors and whether stronger regulations on potential pollutants could reduce mortality.

Results

In order to account for the confounding variables in the model, we started with trying to find a basic model containing only significant weather and socioeconomic factors. A basic exploratory data analysis such as checking correlation between predictors with a scatterplot matrix was conducted. To achieve linearity, log transformations were implemented on terms NonWhite, Density, and Poor. Meanwhile, case 7, which is Miami, FL, and case 20, which is York, PA were removed as influential cases, as they influenced the significance level in some variables. Miami has the highest precipitation and a low mortality, causing it to have the highest cook's distance. York has the highest density and the lowest education level, which affected the significance level of the education coefficient. After comparing the nested models with full models using anova, insignificant terms were all removed, and the resulting model of mortality against weather and demographic factors is:

$$\hat{\mu}\{Mortality \mid weather, demographics\} = 517.076 + 2.475 Precip - 20.105 Educ + 29.316 log(NonWhite) \\ + 60.112 log(Density)$$

After checking the assumptions and the outliers of the model above, the variance inflation factor of each term is not high enough for us to remove any possible collinear parameter.

Since the base model is good enough, we started to add pollution variables into the model. Using a scatterplot matrix, we logged all three of the pollution variables to achieve linearity. After checking model assumptions and partial residual plots, we checked the outliers of the model, and case 60, which is New Orleans, LA, is deemed as an influential outlier as it has the highest mortality but the lowest amount of SO2. Additionally, case 4, which is Lancaster, PA, is also deemed as an influential outlier as it has one of the lowest education levels and an unusually low mortality. After checking the model's assumptions, partial residual plots, and collinearity, insignificant terms were removed from the model and the final model of mortality against pollution accounting for weather and demographic factors is shown below.

$$\begin{aligned}\hat{\mu}\{Mortality \mid weather, demographics, pollution\} &= 692.490 + 2.525 Precip - 16.576 Educ \\ &+ 22.312 log(NonWhite) + 30.404 log(Density) \\ &+ 12.918 log(SO2)\end{aligned}$$

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	692.490	118.156	5.861	0.000	455.167	929.813
Precip	2.525	0.436	5.794	0.000	1.650	3.400
Educ	-16.576	5.471	-3.030	0.004	-27.566	-5.587
log(NonWhite)	22.312	4.101	5.440	0.000	14.074	30.550
log(Density)	30.404	12.422	2.448	0.018	5.454	55.353
log(SO2)	12.918	3.235	3.993	0.000	6.419	19.416

Discussion

According to the regression model, potential pollution as measured by log of SO2 is associated with mortality after accounting for weather and socioeconomic factors. The relative pollution potential of sulfur dioxide has the strongest association with mortality after accounting for weather and demographic factors, while the relative pollution potential of hydrocarbons and oxides of nitrogen don't have a statistically significant association with mortality. One unit increase in relative pollution potential of sulfur dioxide is associated with an increase of 12.918 units in total age-adjusted mortality from all causes holding other variables constant. Our model suggests that stronger regulations on potential pollutants could reduce mortality.

One of the limitations of our model is that the data lack independence due to spatial correlation. The data set is also relatively small with a size of 60, therefore the smooth line would be influenced heavily by any outlier. Some cases don't even have a Cook's distance above one, but removing them still causes a change in the significance levels of variables. This might be the result of a small dataset. Even though both the r squared value and adjusted r squared value are above 80%, there still might be better models that are not in our knowledge to find the association between potential pollution and mortality.

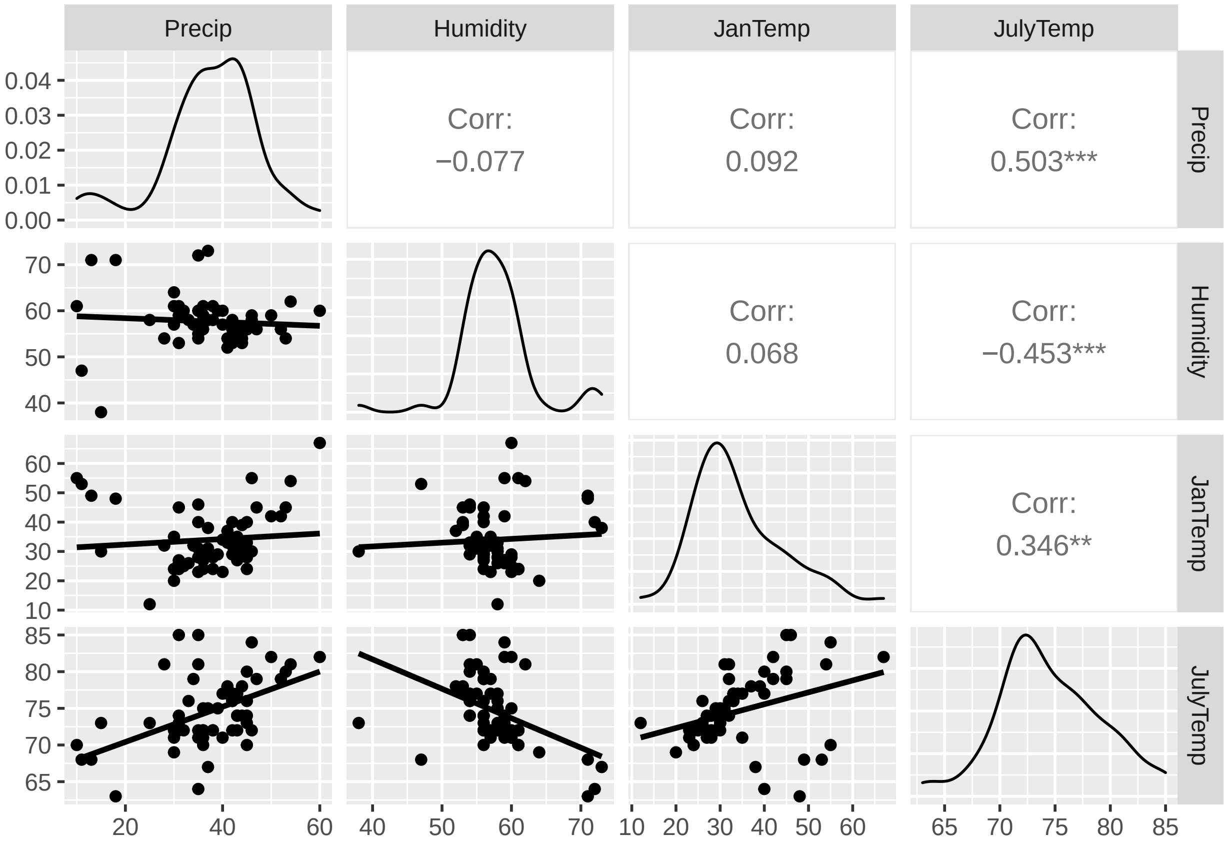
R Code Appendix

```
# import libraries
library(Sleuth3)
library(ggplot2)
library(skimr)
library(ggResidpanel)
library(car)
library(GGally)
library(dplyr)
library(knitr)
library(broom)
```

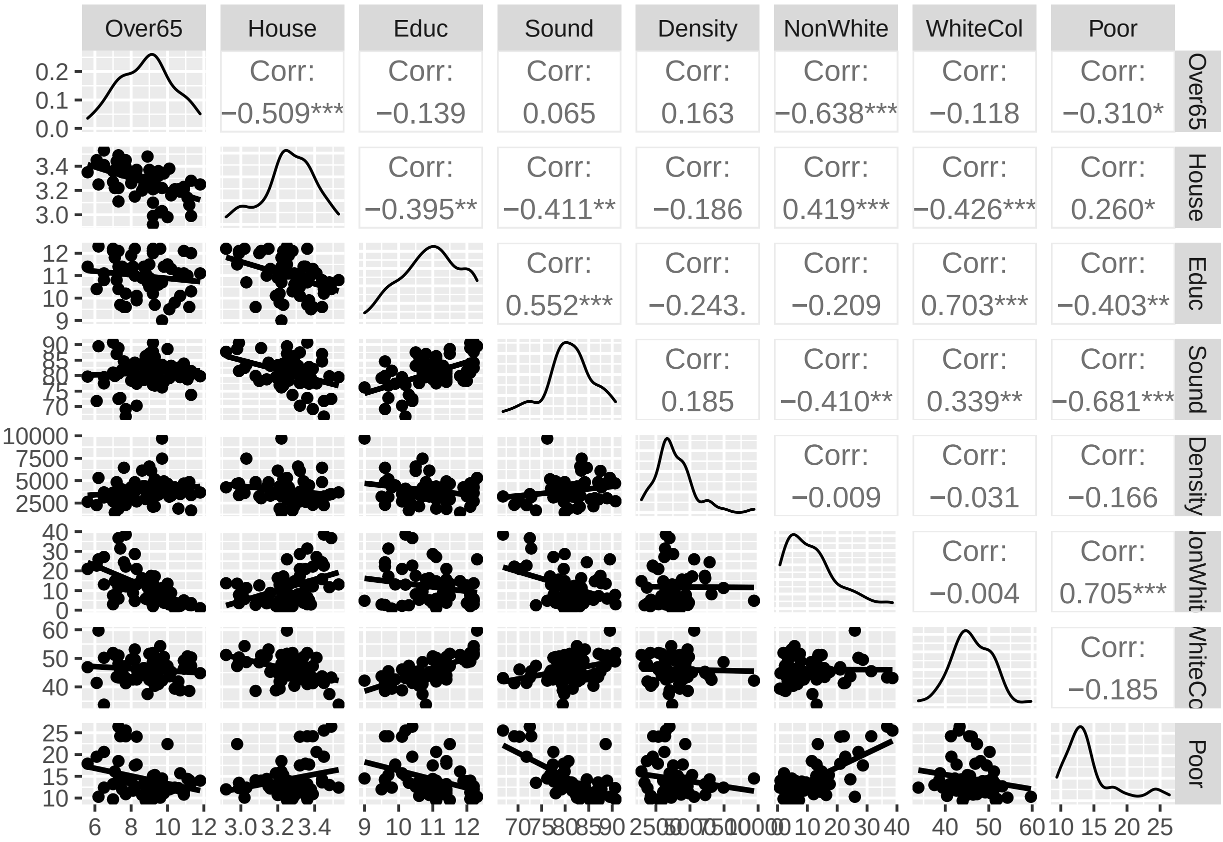
```
# glance on the data and model
pm <- ex1217
summary(pm)
```

```
##             CITY       Mortality         Precip         Humidity
##  Akron, OH     : 1   Min.   : 790.7   Min.   :10.00   Min.   :38.00
##  Albany, NY    : 1   1st Qu.: 898.4   1st Qu.:32.75   1st Qu.:55.00
##  Allentown, PA : 1   Median : 943.7   Median :38.00   Median :57.00
##  Atlanta, GA   : 1   Mean   : 940.4   Mean   :37.37   Mean   :57.67
##  Baltimore, MD : 1   3rd Qu.: 983.2   3rd Qu.:43.25   3rd Qu.:60.00
##  Birmingham, AL: 1   Max.   :1113.1   Max.   :60.00   Max.   :73.00
##  (Other)       :54
##     JanTemp         JulyTemp         Over65           House
##  Min.   :12.00   Min.   :63.00   Min.   : 5.600   Min.   :2.920
##  1st Qu.:27.00   1st Qu.:72.00   1st Qu.: 7.675   1st Qu.:3.210
##  Median :31.50   Median :74.00   Median : 9.000   Median :3.265
##  Mean   :33.98   Mean   :74.58   Mean   : 8.798   Mean   :3.263
##  3rd Qu.:40.00   3rd Qu.:77.25   3rd Qu.: 9.700   3rd Qu.:3.360
##  Max.   :67.00   Max.   :85.00   Max.   :11.800   Max.   :3.530
##
##       Educ           Sound          Density        NonWhite        WhiteCol
##  Min.   : 9.00   Min.   :66.80   Min.   :1441   Min.   : 0.80   Min.   :33.80
##  1st Qu.:10.40   1st Qu.:78.38   1st Qu.:3104   1st Qu.: 4.95   1st Qu.:43.25
##  Median :11.05   Median :81.15   Median :3567   Median :10.40   Median :45.50
##  Mean   :10.97   Mean   :80.91   Mean   :3875   Mean   :11.87   Mean   :46.08
##  3rd Qu.:11.50   3rd Qu.:83.60   3rd Qu.:4520   3rd Qu.:15.65   3rd Qu.:49.52
##  Max.   :12.30   Max.   :90.70   Max.   :9699   Max.   :38.50   Max.   :59.70
##
##       Poor             HC              NOX              SO2
##  Min.   : 9.40   Min.   :  1.00   Min.   :  1.00   Min.   :  1.00
##  1st Qu.:12.00   1st Qu.:  7.00   1st Qu.:  4.00   1st Qu.: 11.00
##  Median :13.20   Median : 14.50   Median :  9.00   Median : 30.00
##  Mean   :14.37   Mean   : 37.85   Mean   : 22.65   Mean   : 53.77
##  3rd Qu.:15.15   3rd Qu.: 30.25   3rd Qu.: 23.75   3rd Qu.: 69.00
##  Max.   :26.40   Max.   :648.00   Max.   :319.00   Max.   :278.00
##
```

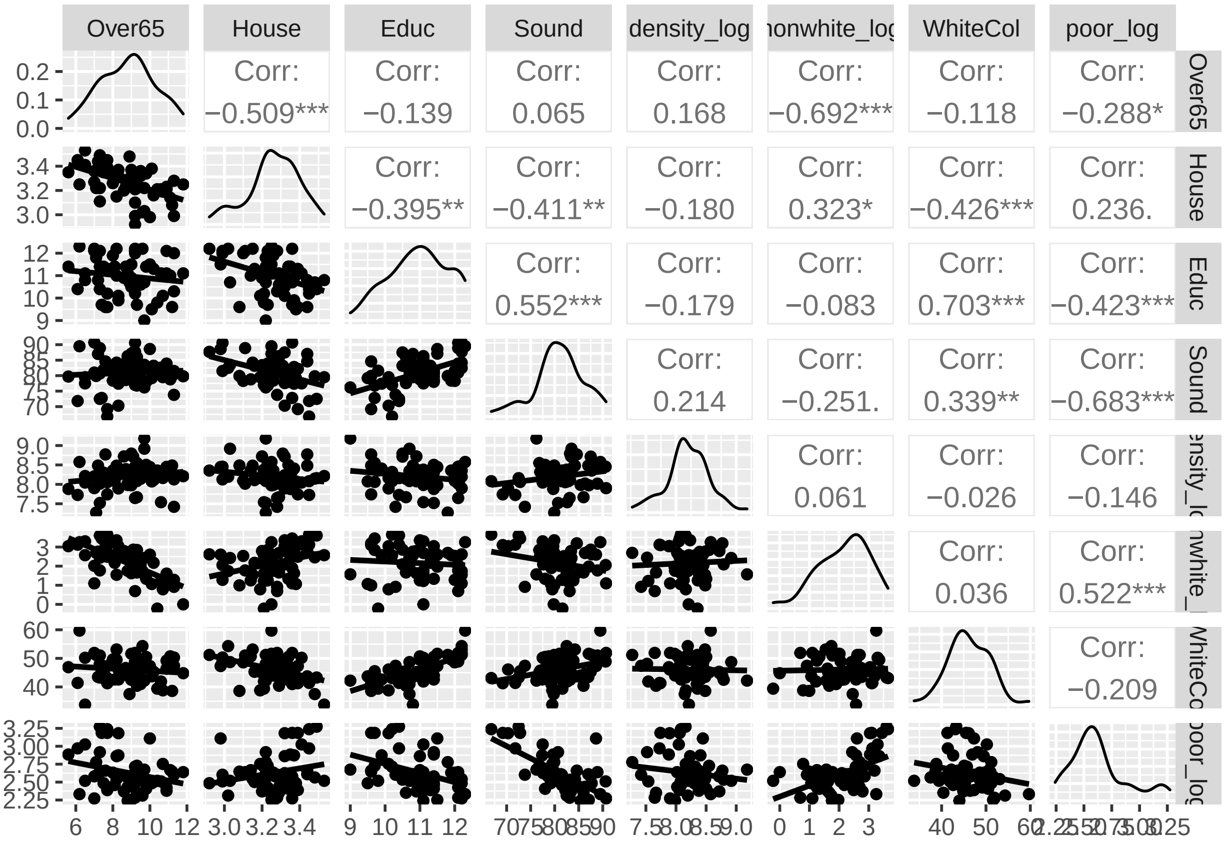
```
# ggpairs - weather subset
ggpairs(pm,
        columns = c("Precip", "Humidity", "JanTemp", "JulyTemp"),
        lower = list(continuous =  wrap("smooth", se = FALSE)))
```



```
# ggpairs - demographics subset
ggpairs(pm,
       columns = c("Over65", "House", "Educ", "Sound", "Density", "NonWhite", "WhiteCol", "Poor"),
       lower = list(continuous =  wrap("smooth", se = FALSE)))
```



```
# ggpairs - transformed demographics subset
pm %>%
  mutate(
    density_log = log(Density),
    nonwhite_log = log(NonWhite),
    poor_log = log(Poor)
  ) %>%
  ggpairs(columns = c("Over65", "House", "Educ", "Sound", "density_log", "nonwhite_log", "WhiteCol", "po
        lower = list(continuous =  wrap("smooth", se = FALSE)),
        columnLabels = c("Over65", "House", "Educ", "Sound", "density_log", "nonwhite_log", "WhiteCol",
```

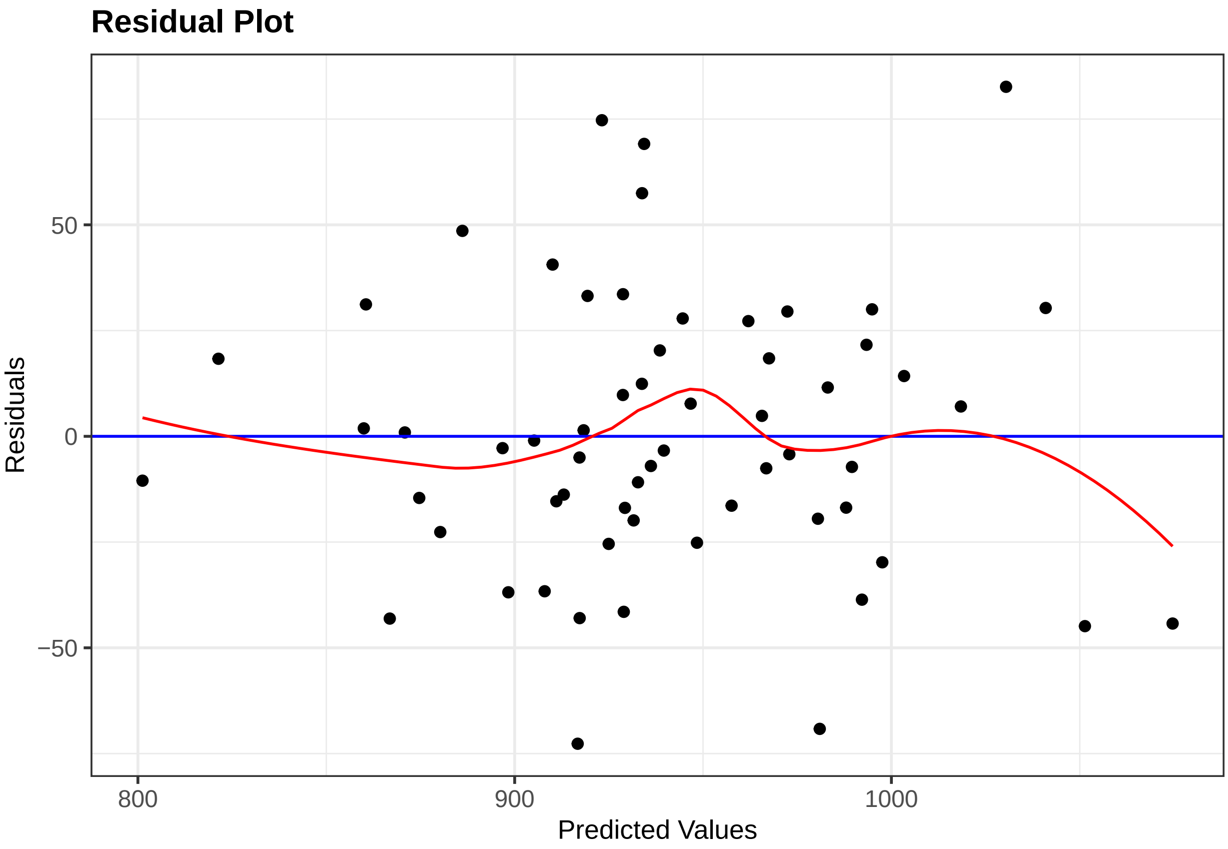


```
# untranformed model - weather and demographics
weather_demo_unlog_lm <- lm(Mortality~Precip+Humidity+JanTemp+JulyTemp+Over65+House+Educ+Sound+Density+
summary(weather_demo_unlog_lm)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Humidity + JanTemp + JulyTemp +
##     Over65 + House + Educ + Sound + Density + NonWhite + WhiteCol +
##     Poor, data = pm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.677 -19.583  -3.084  20.636  82.627
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.770e+03  4.443e+02   3.984 0.000234 ***
## Precip       1.572e+00  8.250e-01   1.906 0.062842 .
## Humidity    -1.145e-01  1.104e+00  -0.104 0.917840
## JanTemp     -2.166e+00  9.995e-01  -2.167 0.035349 *
## JulyTemp    -3.103e+00  1.859e+00  -1.669 0.101750
## Over65      -4.593e+00  8.267e+00  -0.556 0.581169
## House       -1.033e+02  7.238e+01  -1.428 0.160027
## Educ        -2.089e+01  1.122e+01  -1.861 0.068970 .
## Sound       -3.761e-01  1.814e+00  -0.207 0.836618
## Density      5.325e-03  4.174e-03   1.276 0.208298
## NonWhite     5.741e+00  1.157e+00   4.962 9.58e-06 ***
## WhiteCol    -3.992e-01  1.644e+00  -0.243 0.809197
## Poor        -7.119e-01  3.291e+00  -0.216 0.829669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.68 on 47 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.6523
## F-statistic: 10.22 on 12 and 47 DF,  p-value: 1.829e-09
```

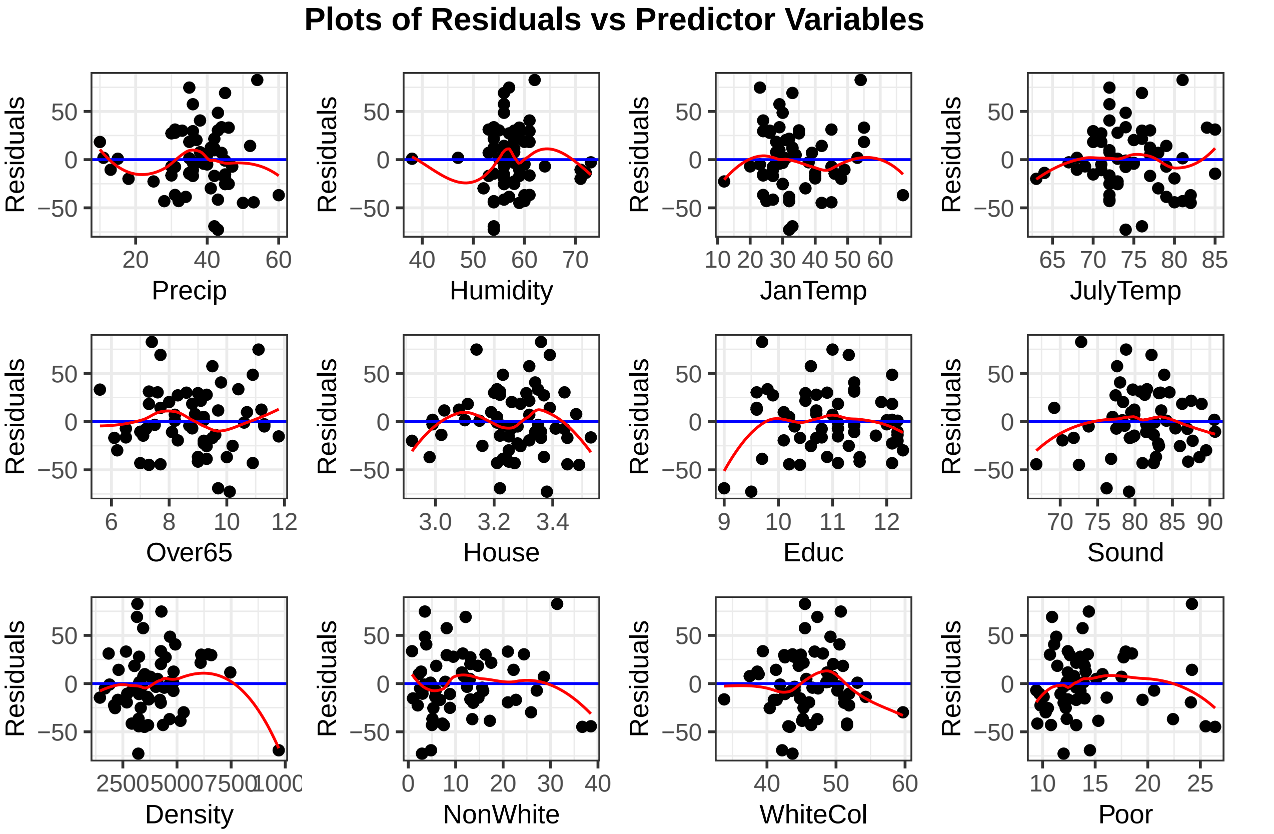
```
# check assumptions 1 - weather and demographics
# residuals plot
resid_panel(weather_demo_unlog_lm, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

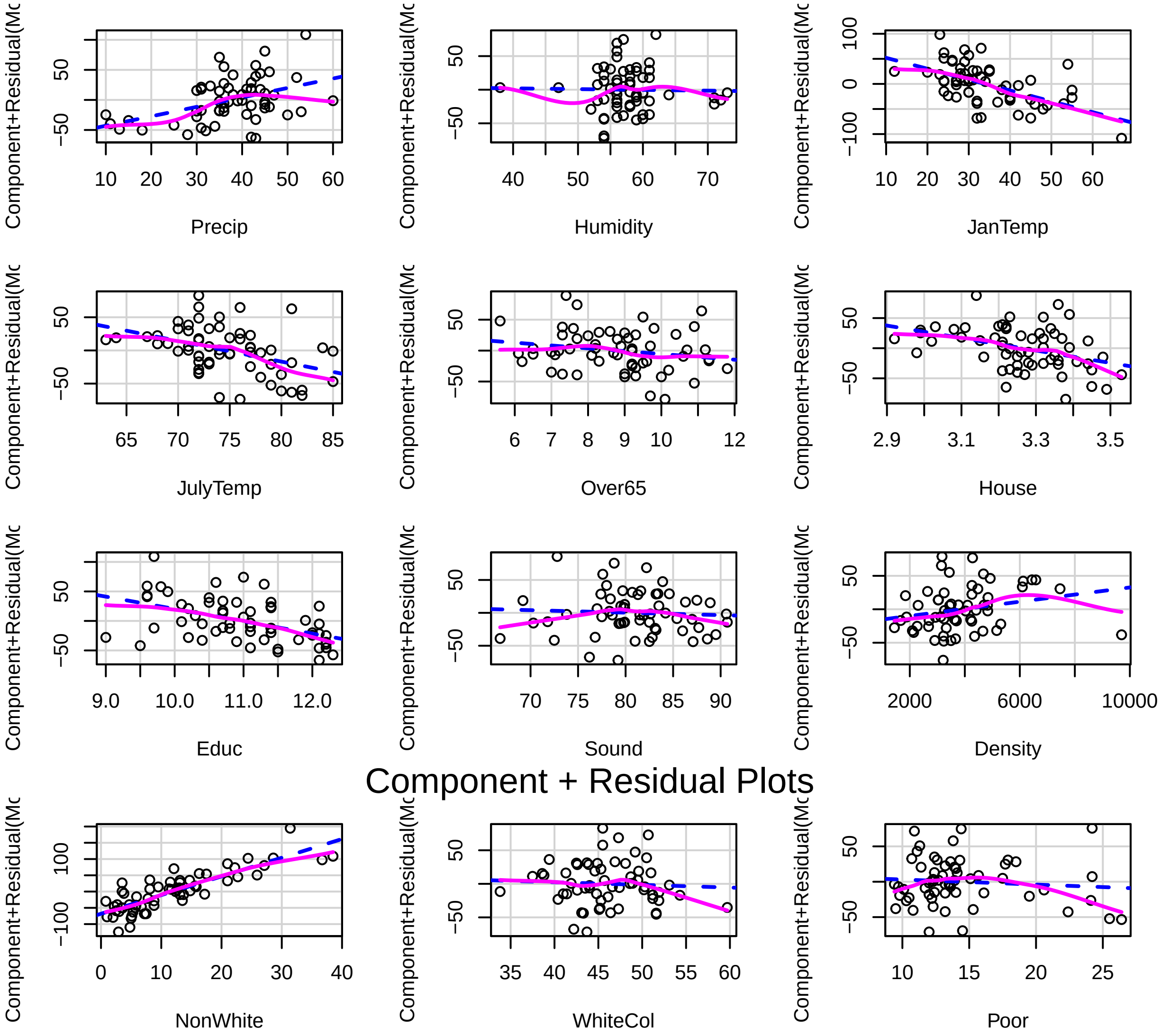


```
# residuals of each predictor
resid_xpanel(weather_demo_unlog_lm, smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



```r
# partial residuals
crp(weather_demo_unlog_lm)
```

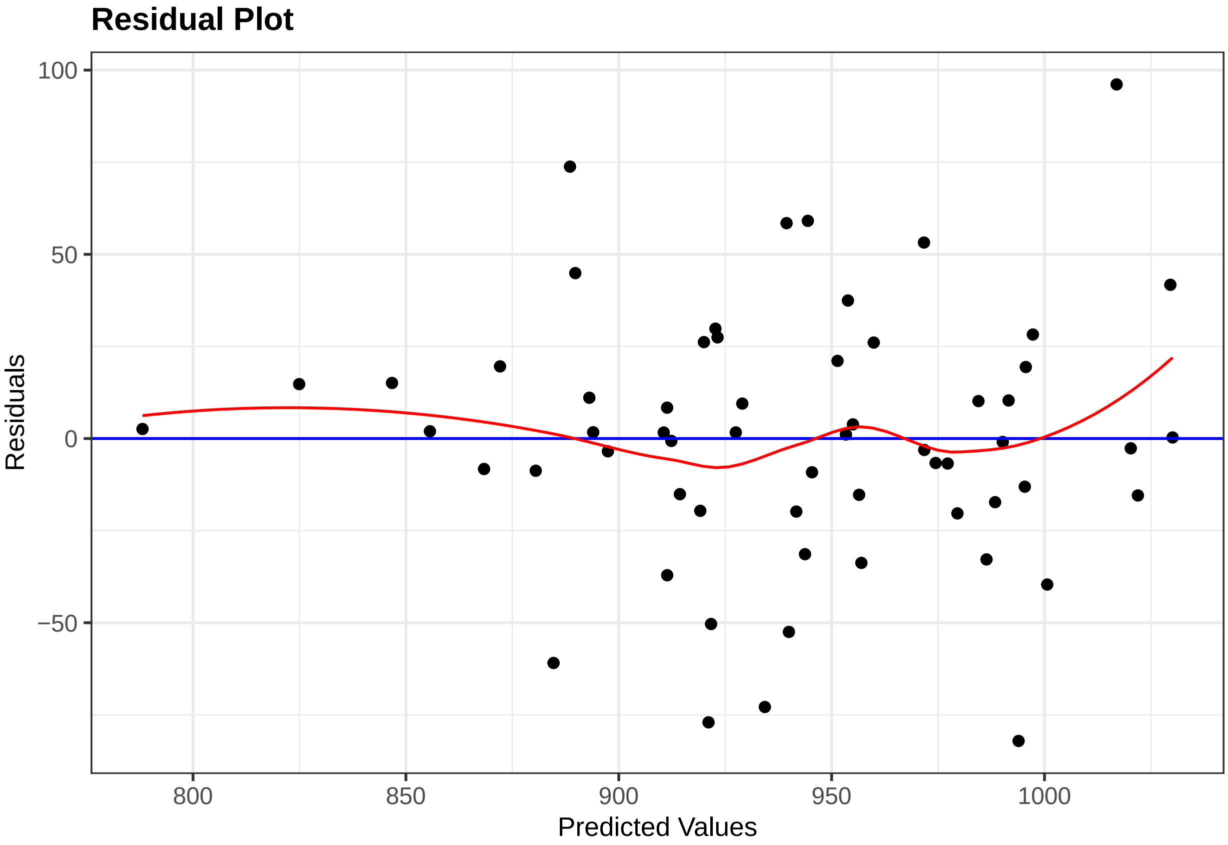


```
# transformed model - weather and demographics
weather_demo_log_lm <- lm(Mortality~Precip+Humidity+JanTemp+JulyTemp+Over65+House+Educ+Sound+log(Densit
summary(weather_demo_log_lm)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Humidity + JanTemp + JulyTemp +
##     Over65 + House + Educ + Sound + log(Density) + log(NonWhite) +
##     WhiteCol + log(Poor), data = pm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.089 -15.901   0.689  19.461  96.126
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1189.4976   528.3295   2.251  0.02907 *
## Precip           2.4265     0.8391   2.892  0.00579 **
## Humidity        -0.2146     1.1895  -0.180  0.85759
## JanTemp         -2.4964     1.0712  -2.331  0.02412 *
## JulyTemp        -4.1670     1.9977  -2.086  0.04244 *
## Over65          -8.5824     9.1121  -0.942  0.35108
## House          -57.3601    77.2313  -0.743  0.46136
## Educ           -23.4524    11.3488  -2.067  0.04432 *
## Sound            0.2522     1.9650   0.128  0.89840
## log(Density)    30.2566    17.2977   1.749  0.08679 .
## log(NonWhite)   37.0958    10.5920   3.502  0.00102 **
## WhiteCol         1.4232     1.7133   0.831  0.41034
## log(Poor)       66.1741    52.5217   1.260  0.21391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.98 on 47 degrees of freedom
## Multiple R-squared:  0.6872, Adjusted R-squared:  0.6073
## F-statistic: 8.603 on 12 and 47 DF,  p-value: 2.516e-08
```

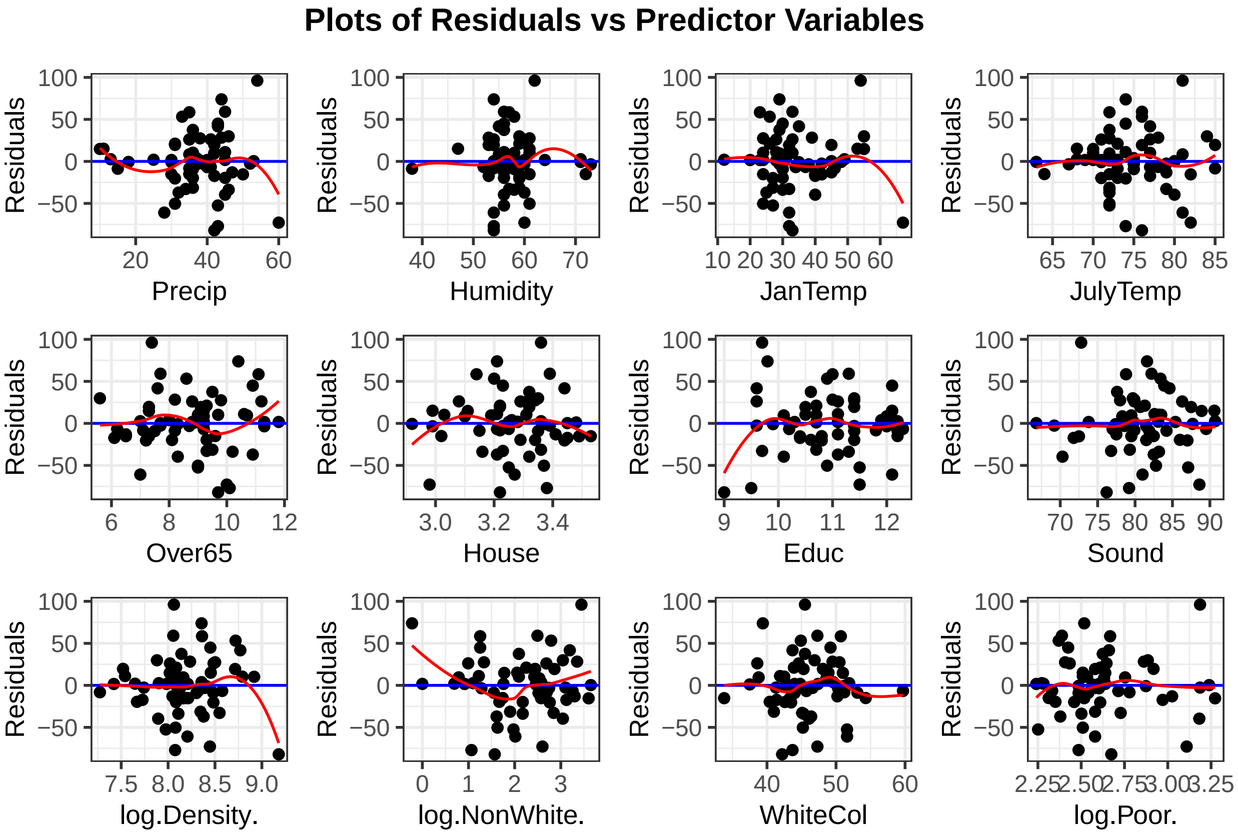
```
# check assumptions 2 - weather and demographics
# residuals plot
resid_panel(weather_demo_log_lm, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

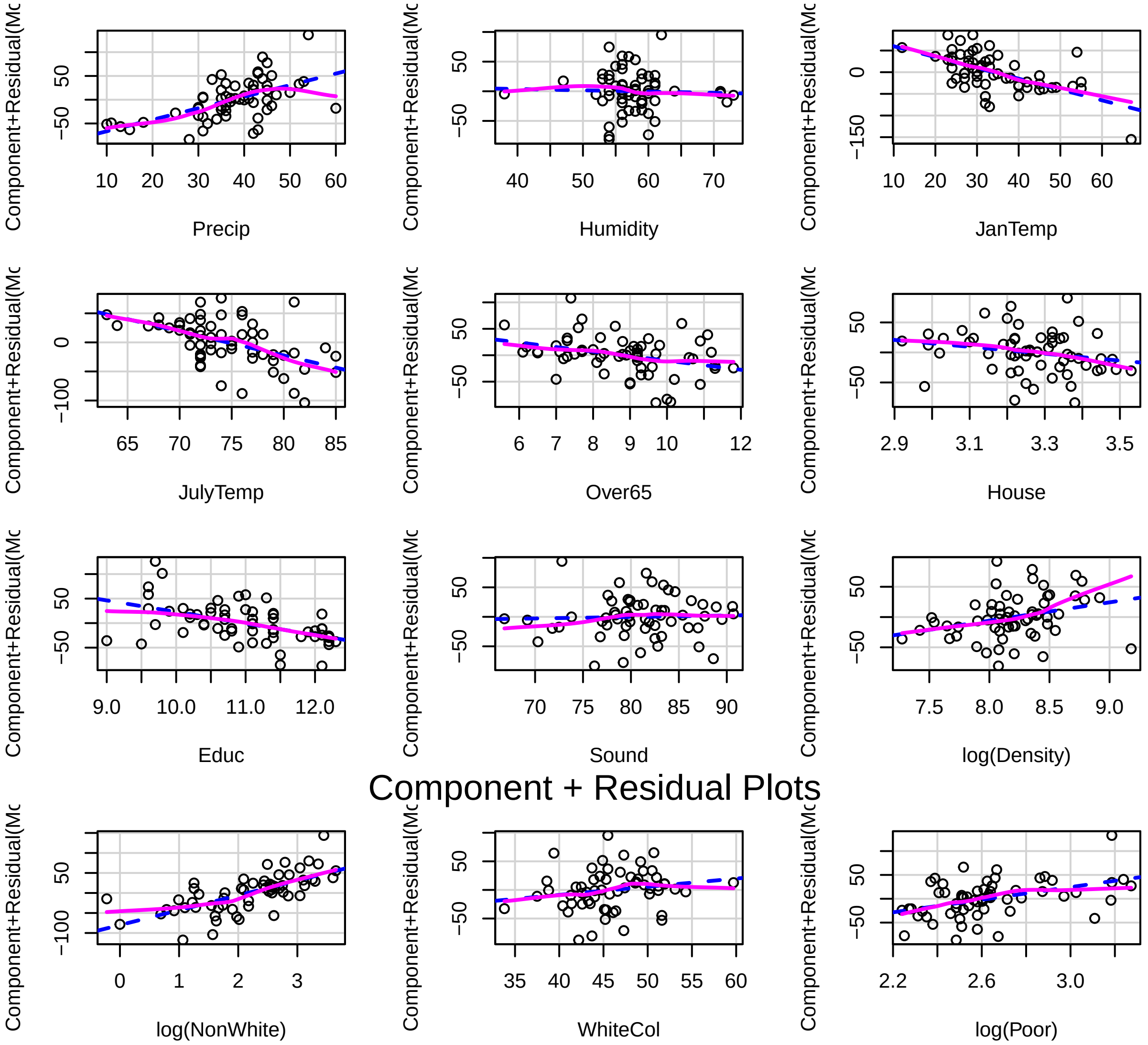


```
# residuals of each predictor
resid_xpanel(weather_demo_log_lm, smoother = TRUE)
```

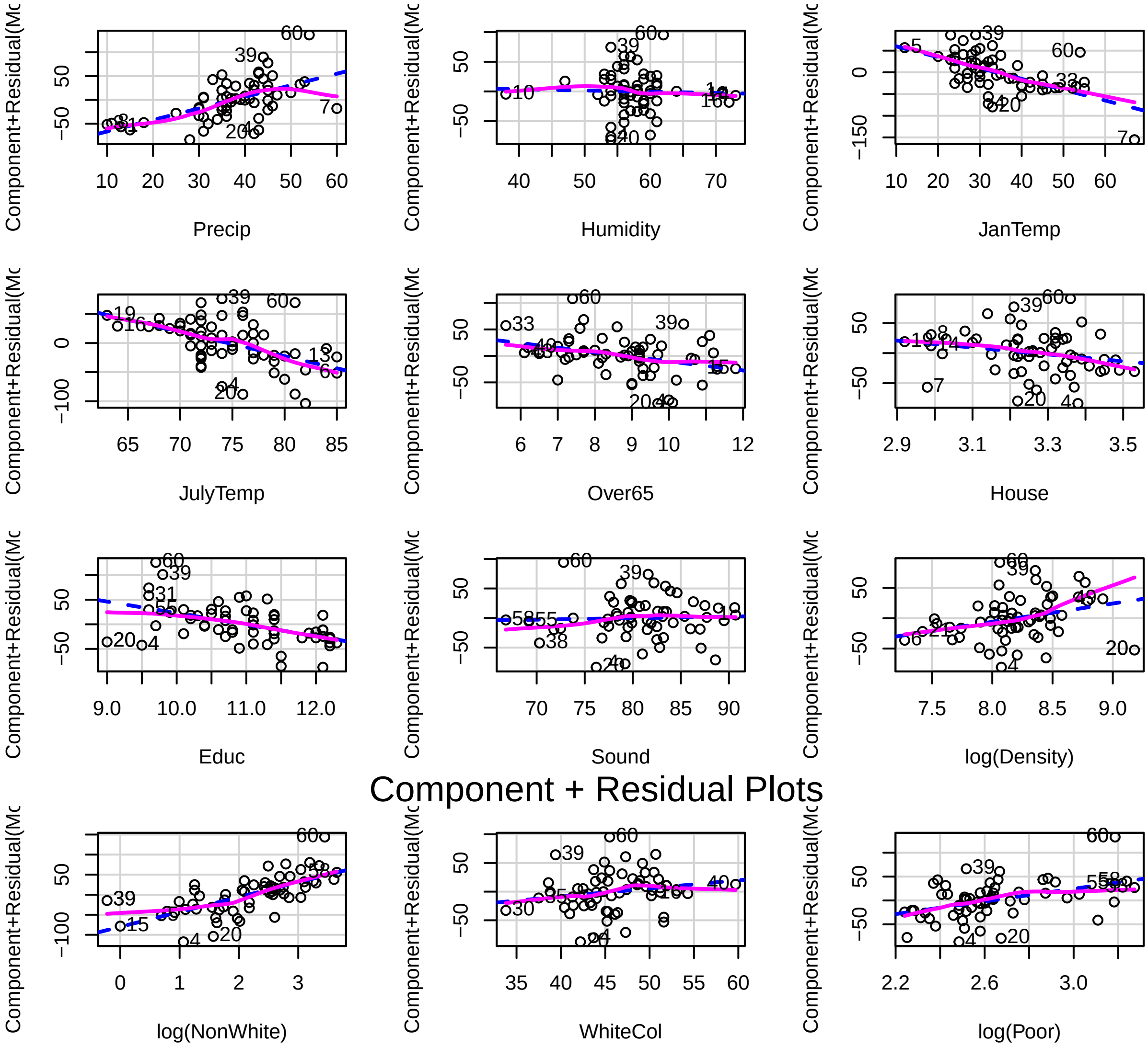
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



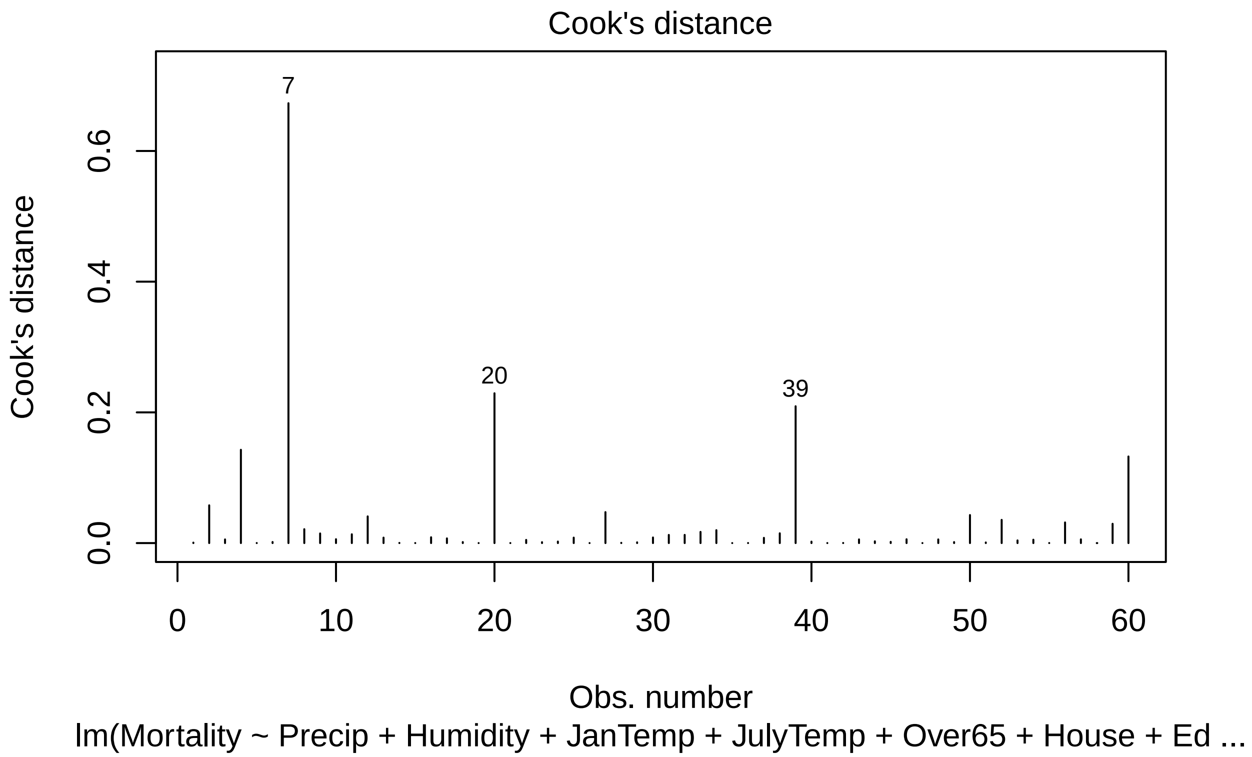
```
# partial residuals
crp(weather_demo_log_lm)
```

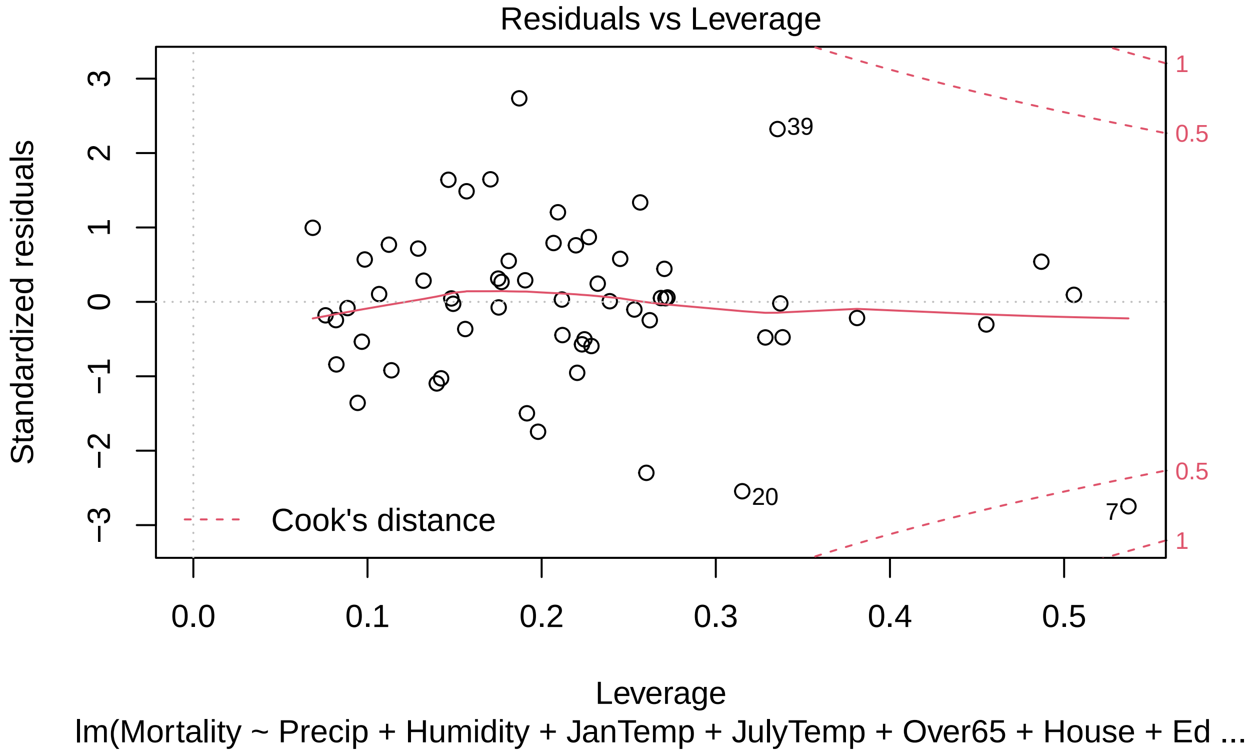


```r
crp(weather_demo_log_lm, id = list(n = 4))
```



```
# check outliers 1 - weather and demographics
plot(weather_demo_log_lm, which =c(4,5))
```





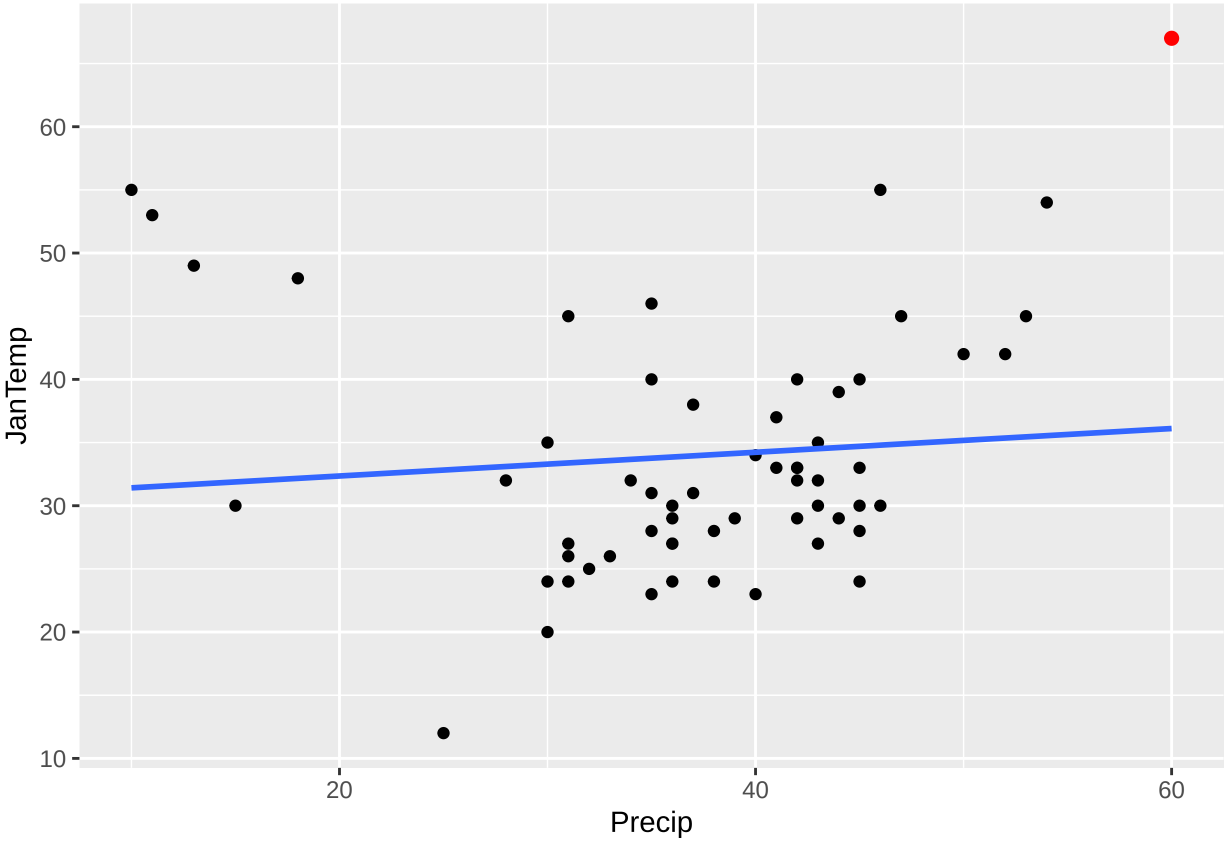
```r
# add case numbers onto the data set
pm_mutate <- pm %>% mutate(case = row_number())

# slice out case 7
pm %>% slice(7)
```

```
##        CITY Mortality Precip Humidity JanTemp JulyTemp Over65 House Educ Sound
## 1 Miami, FL    861.44     60       60      67       82     10  2.98 11.5  88.6
##   Density NonWhite WhiteCol Poor HC NOX SO2
## 1    4657     13.5     47.3 22.4  3   1   1
```

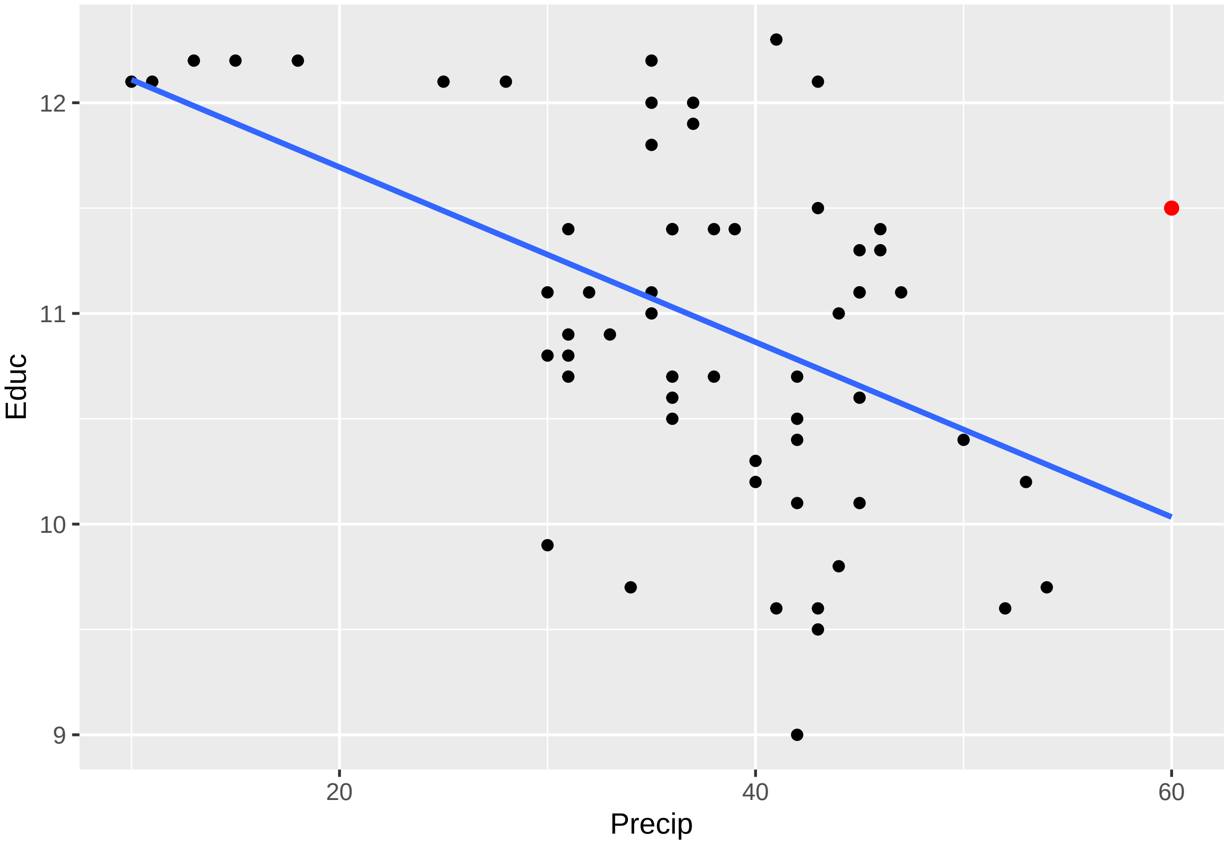
```
# case 7 eda
ggplot(pm_mutate, aes(Precip, JanTemp)) +
  geom_point() +
  geom_point(data=filter(pm_mutate, case == 7), color="red", size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



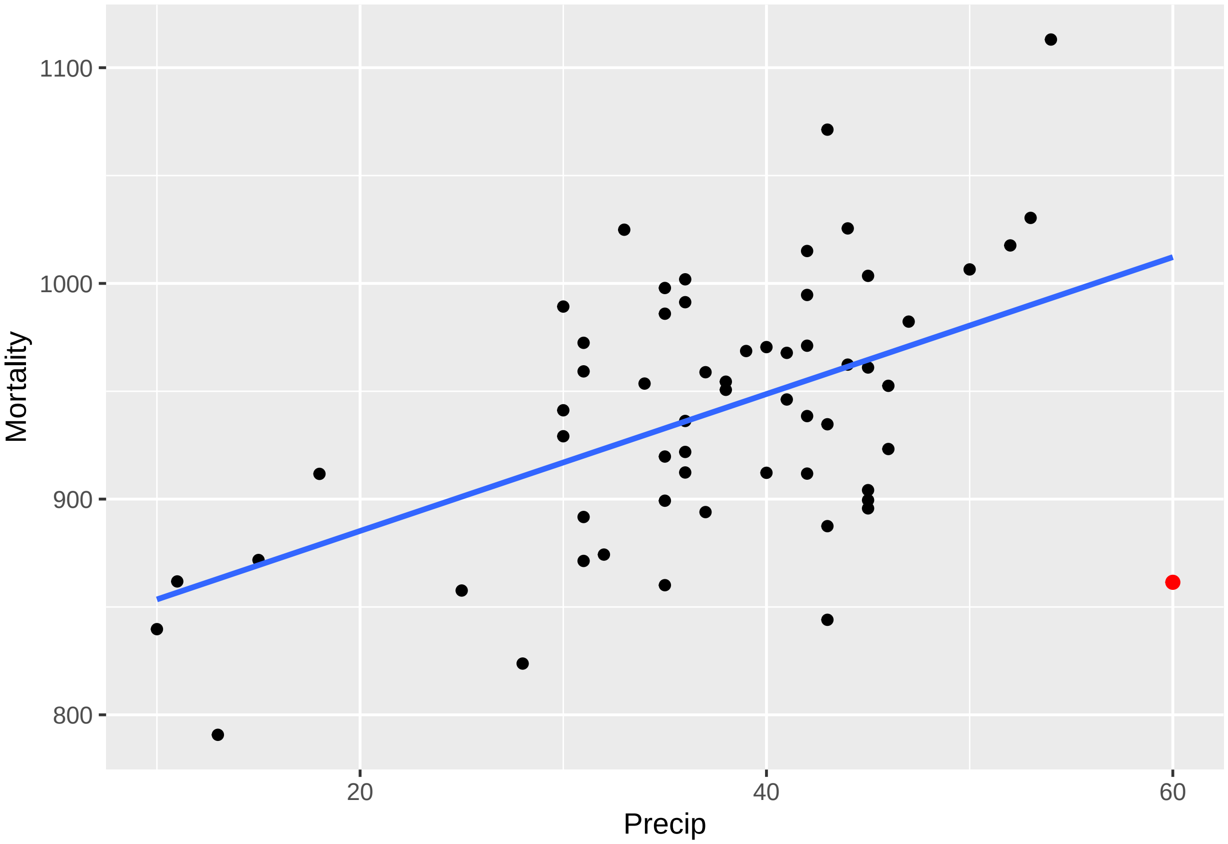
```
ggplot(pm_mutate, aes(Precip, Educ)) +
  geom_point() +
  geom_point(data=filter(pm_mutate, case == 7), color="red", size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(pm_mutate, aes(Precip, Mortality)) +
  geom_point() +
  geom_point(data=filter(pm_mutate, case == 7), color="red", size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

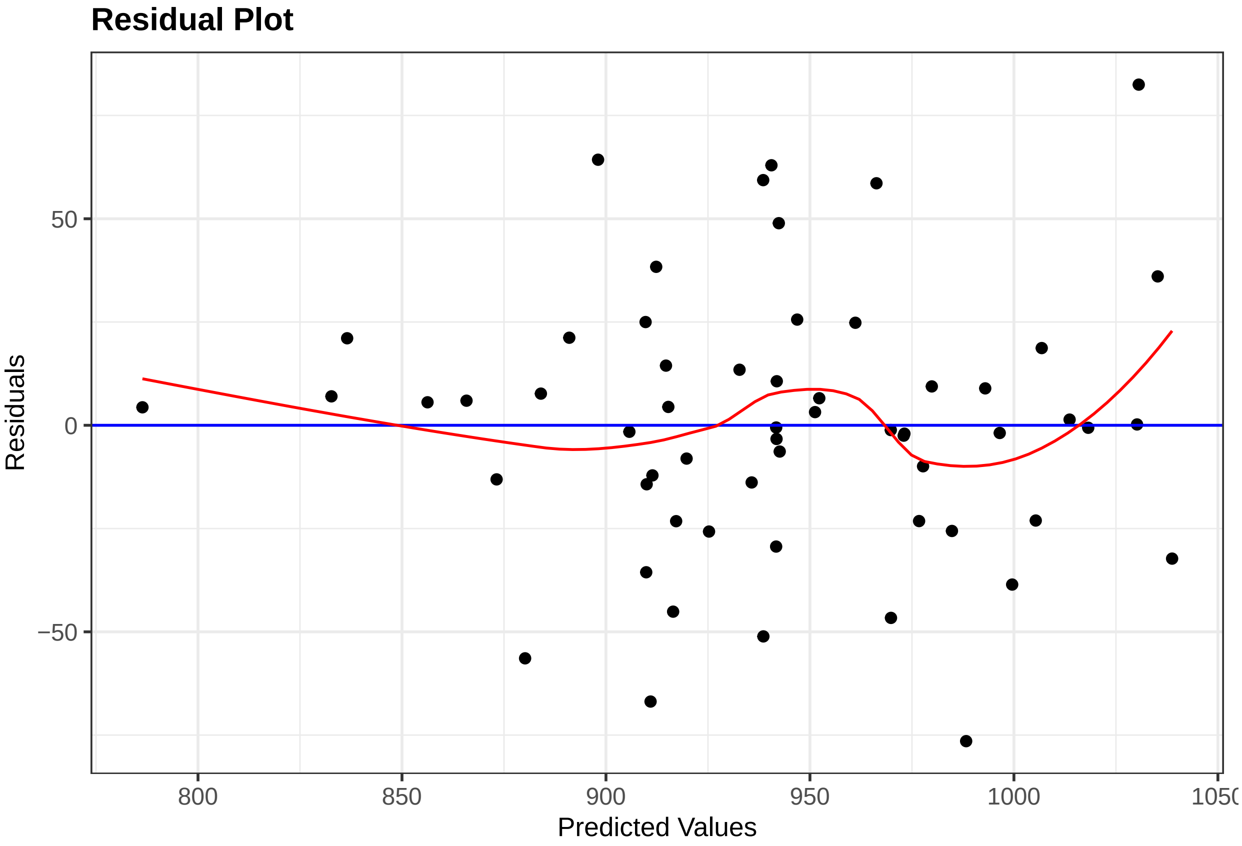


```
# refit model without case 7
weather_demo_log_lm_no_7 <- lm(Mortality~Precip+Humidity+JanTemp+JulyTemp+Over65+House+Educ+Sound+log(De
# Educ is not significant, log(poor) now is significant
summary(weather_demo_log_lm_no_7)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Humidity + JanTemp + JulyTemp +
##     Over65 + House + Educ + Sound + log(Density) + log(NonWhite) +
##     WhiteCol + log(Poor), data = pm, subset = -c(7))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.470 -18.669  -0.545  13.940  82.467
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    632.7829   524.0502   1.207 0.233417
## Precip           3.1091     0.8105   3.836 0.000379 ***
## Humidity         0.2181     1.1112   0.196 0.845294
## JanTemp         -2.1577     0.9986  -2.161 0.035951 *
## JulyTemp        -3.3302     1.8714  -1.779 0.081769 .
## Over65          -5.0279     8.5234  -0.590 0.558149
## House          -58.7501    71.5262  -0.821 0.415667
## Educ           -15.3855    10.8564  -1.417 0.163166
## Sound            2.6820     1.9956   1.344 0.185556
## log(Density)    36.9107    16.1759   2.282 0.027173 *
## log(NonWhite)  34.5599     9.8465   3.510 0.001015 **
## WhiteCol        0.5243     1.6154   0.325 0.746976
## log(Poor)     110.0641    50.8415   2.165 0.035623 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.1 on 46 degrees of freedom
## Multiple R-squared:  0.7299, Adjusted R-squared:  0.6594
## F-statistic: 10.36 on 12 and 46 DF,  p-value: 1.854e-09
```

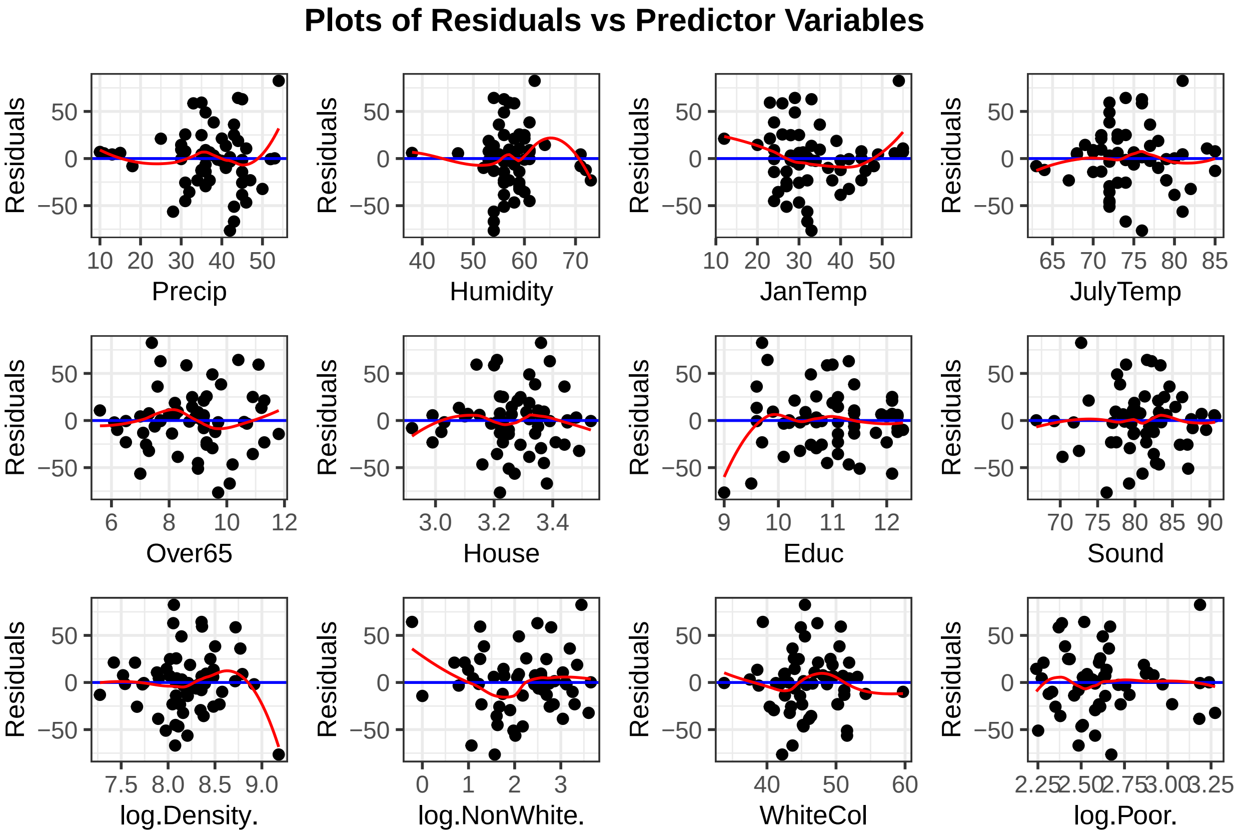
```
# check assumptions 3 - weather and demographics
# residuals plot
resid_panel(weather_demo_log_lm_no_7, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

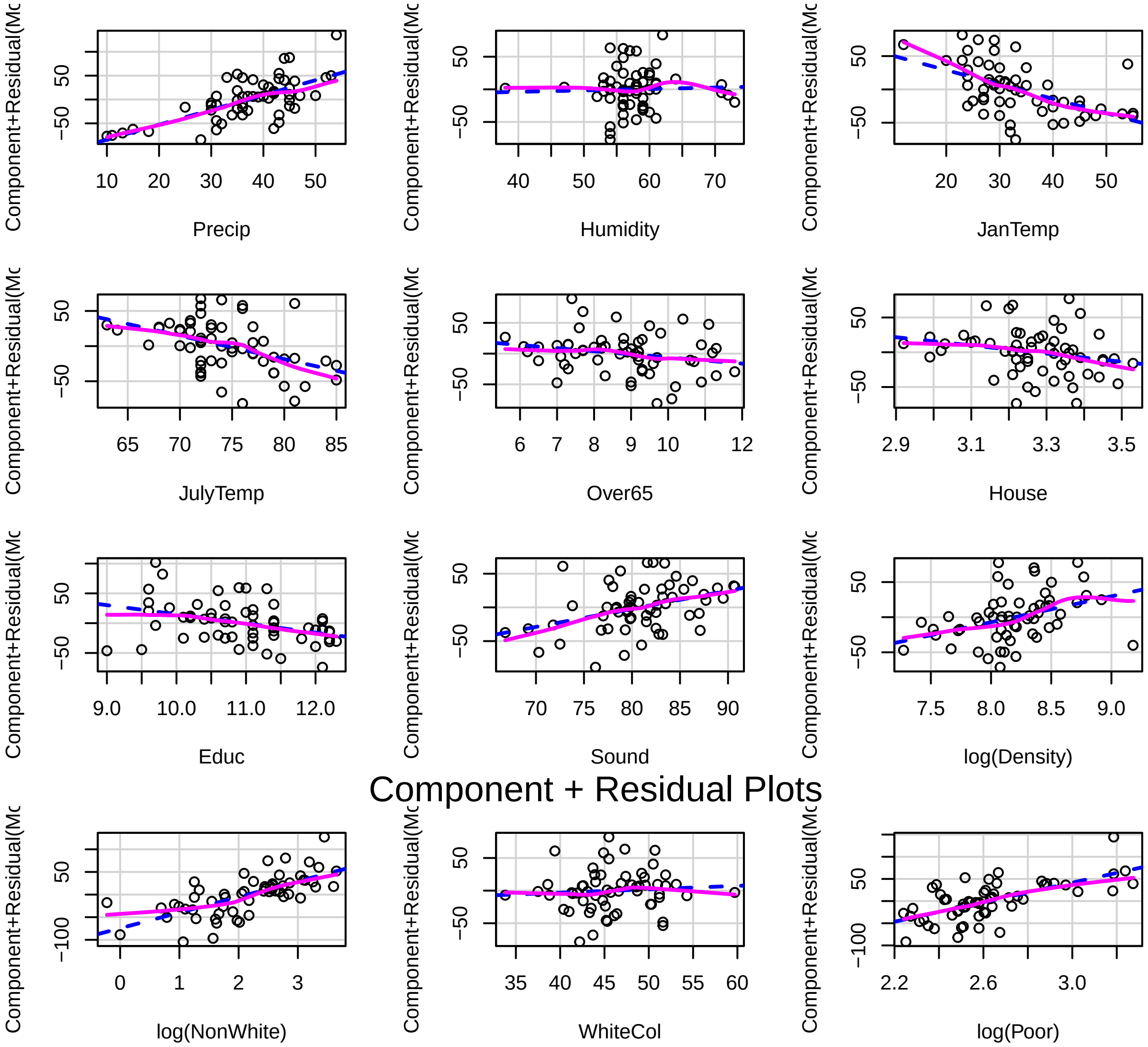


```
# residuals of each predictor
resid_xpanel(weather_demo_log_lm_no_7, smoother = TRUE)
```

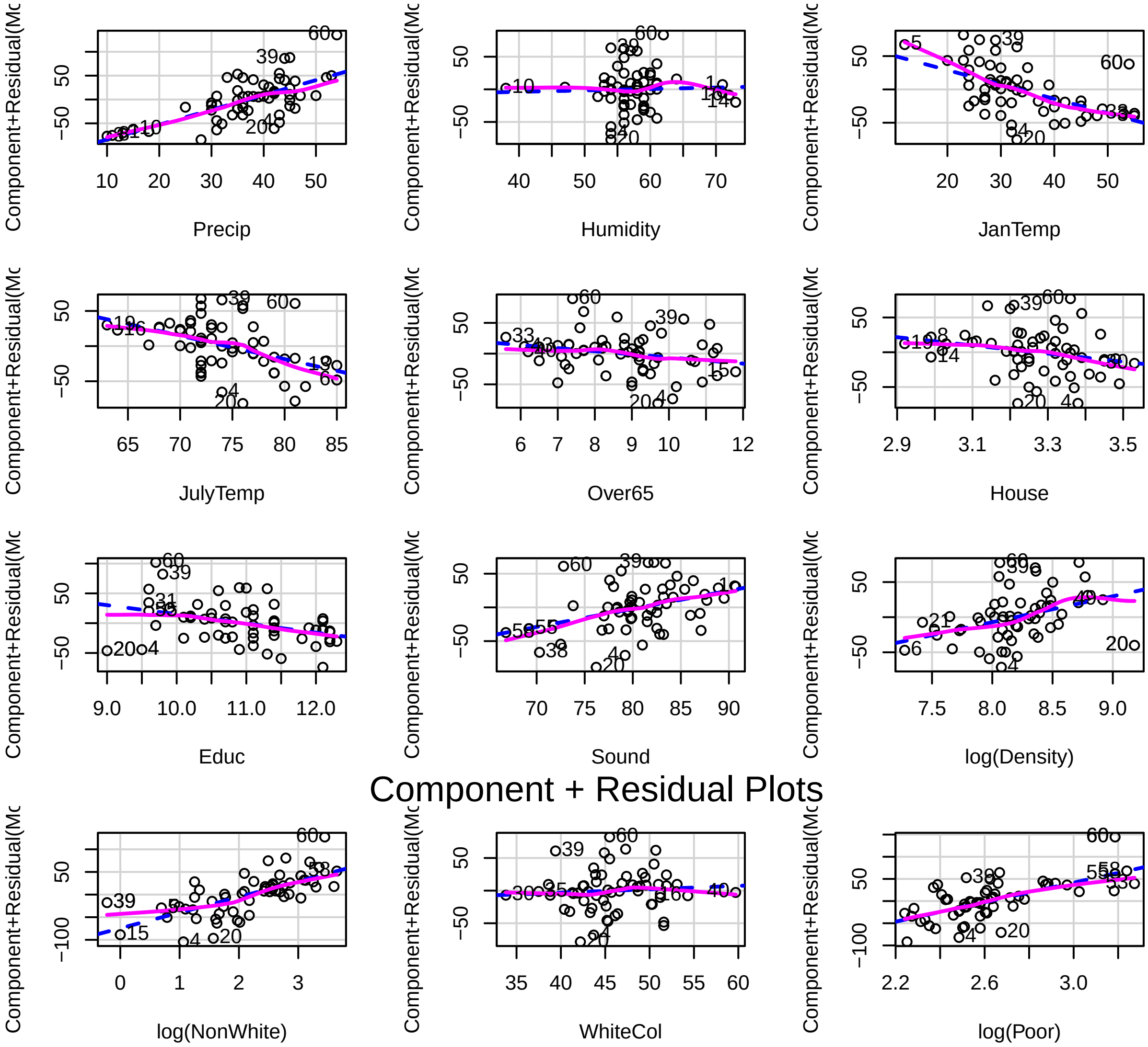
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



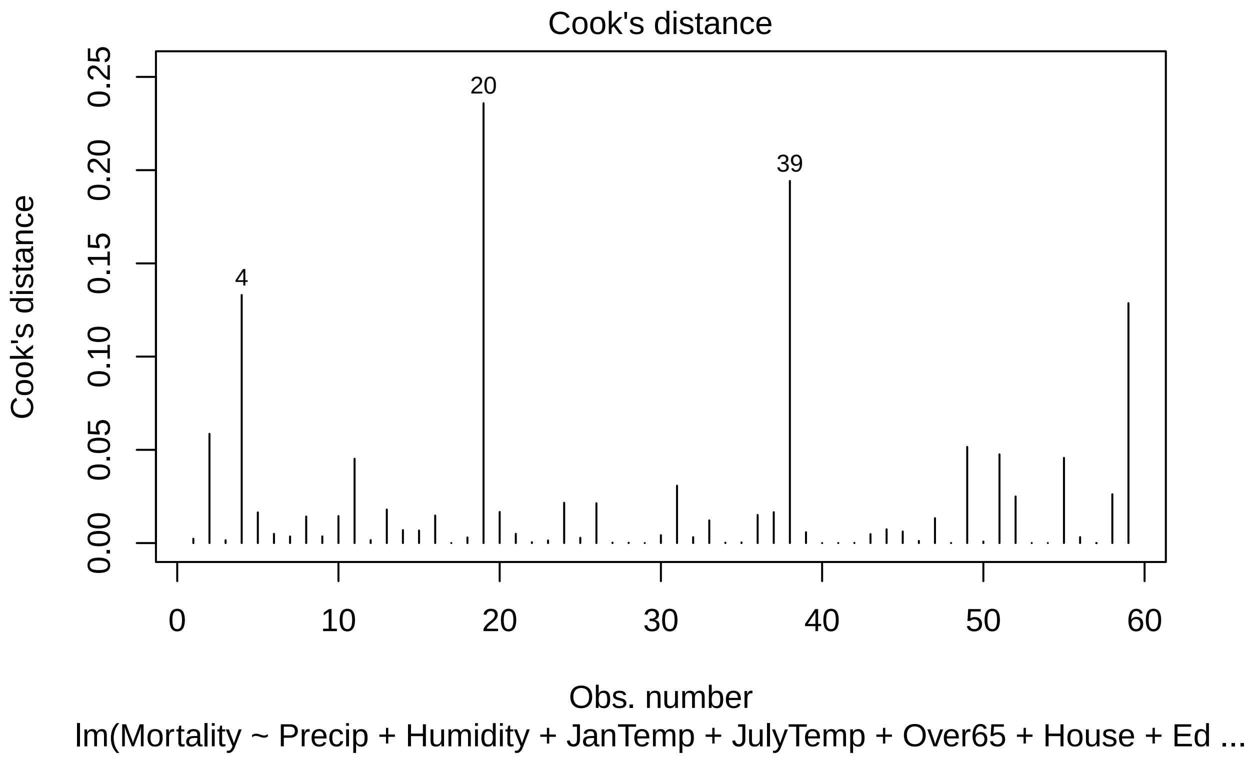
```
# partial residuals
crp(weather_demo_log_lm_no_7)
```

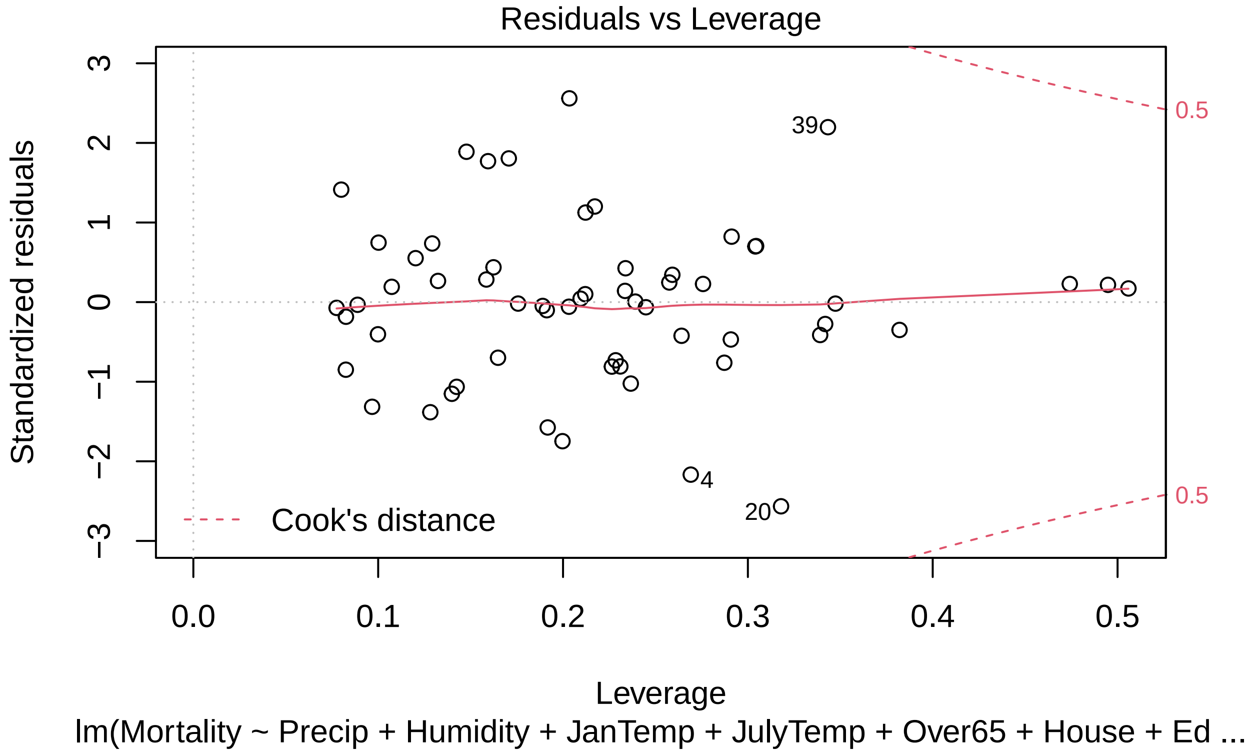


```
crp(weather_demo_log_lm_no_7, id = list(n = 4))
```



```r
# check outliers 2 - weather and demographics
plot(weather_demo_log_lm_no_7, which =c(4,5))
```





```
# refit model without case 7 and case 20
weather_demo_log_lm_no_7_20 <- lm(Mortality~Precip+Humidity+JanTemp+JulyTemp+Over65+House+Educ+Sound+log
# Educ is significant again, JanTemp and log(Poor) are not anymore
summary(weather_demo_log_lm_no_7_20)
```

```
##
```

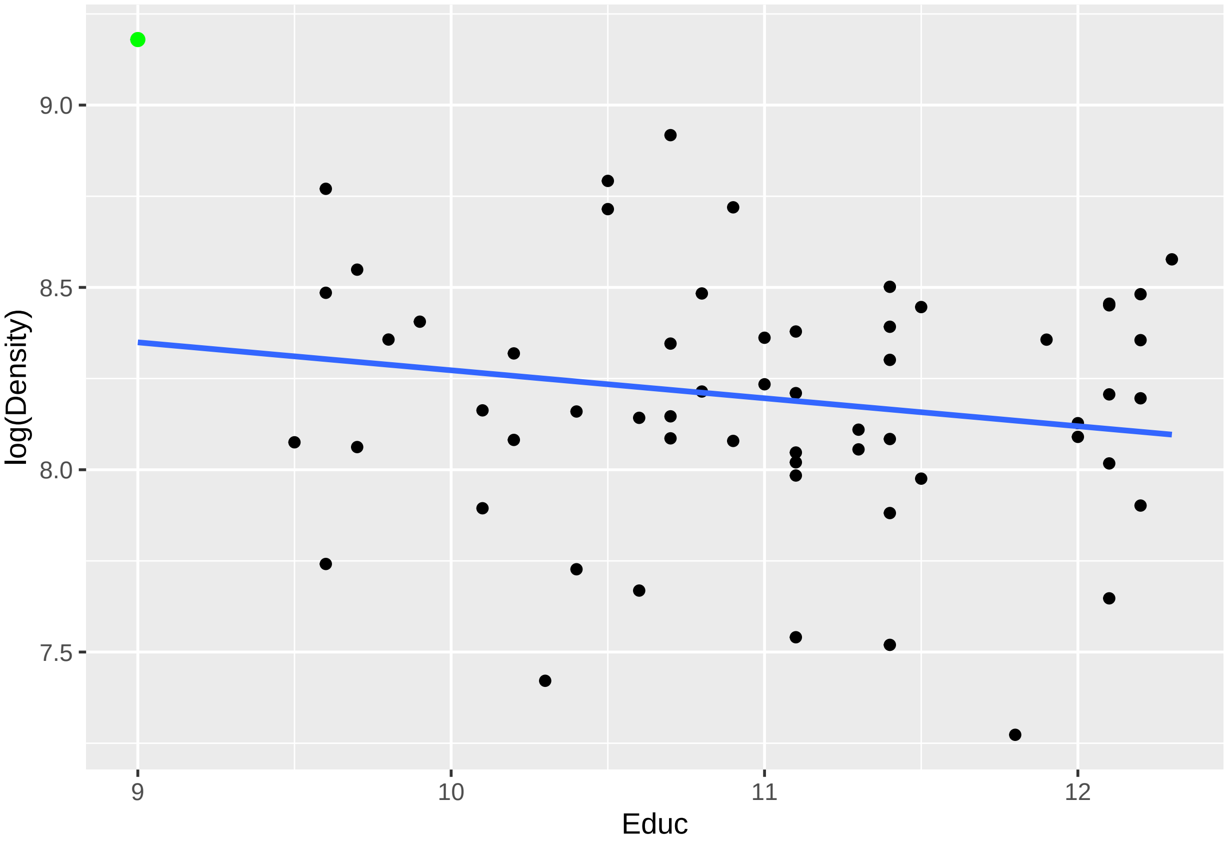
```
## Call:
## lm(formula = Mortality ~ Precip + Humidity + JanTemp + JulyTemp +
##     Over65 + House + Educ + Sound + log(Density) + log(NonWhite) +
##     WhiteCol + log(Poor), data = pm, subset = -c(7, 20))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.526 -16.986  -2.244  16.202  75.313
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   832.5892   495.8801   1.679 0.100081
## Precip          3.0929     0.7586   4.077 0.000183 ***
## Humidity        0.1679     1.0402   0.161 0.872507
## JanTemp        -1.8594     0.9409  -1.976 0.054296 .
## JulyTemp       -3.1357     1.7530  -1.789 0.080391 .
## Over65         -8.6753     8.0879  -1.073 0.289155
## House         -80.8820    67.4311  -1.199 0.236620
## Educ          -22.3448    10.4737  -2.133 0.038377 *
## Sound           1.4538     1.9208   0.757 0.453076
## log(Density)   52.2738    16.1446   3.238 0.002264 **
## log(NonWhite)  28.6651     9.4637   3.029 0.004055 **
## WhiteCol        0.7511     1.5142   0.496 0.622265
## log(Poor)      86.1750    48.3775   1.781 0.081613 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.79 on 45 degrees of freedom
## Multiple R-squared:  0.7676, Adjusted R-squared:  0.7056
## F-statistic: 12.38 on 12 and 45 DF,  p-value: 1.398e-10
```

```
# slice out case 20
pm %>% slice(20)
```

```
##       CITY Mortality Precip Humidity JanTemp JulyTemp Over65 House Educ Sound
## 1 York, PA    911.82     42       54      33       76    9.7  3.22    9  76.2
##   Density NonWhite WhiteCol Poor HC NOX SO2
## 1    9699      4.8     42.2 14.5  8   8  49
```

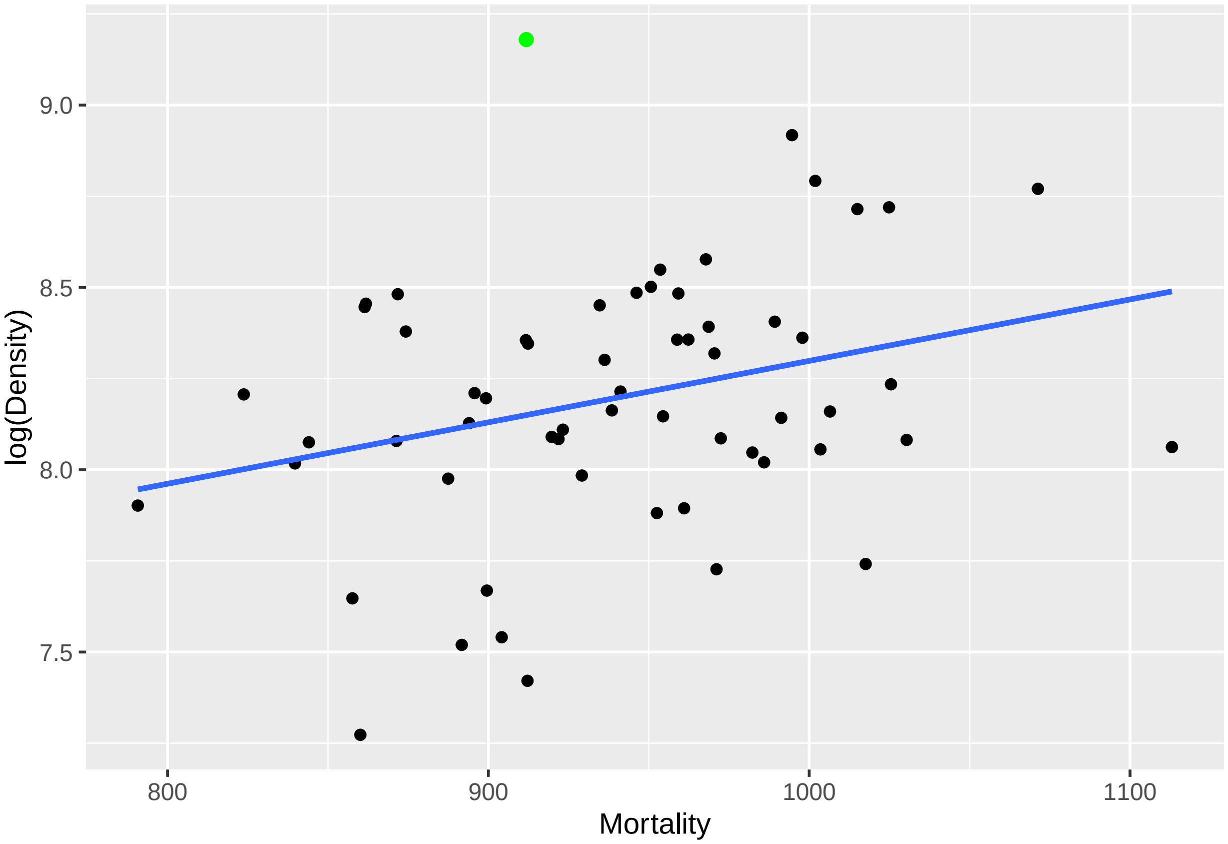
```
# case 20 eda
ggplot(pm_mutate, aes(Educ, log(Density))) +
  geom_point() +
  geom_point(data=filter(pm_mutate, case == 20), color="green", size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

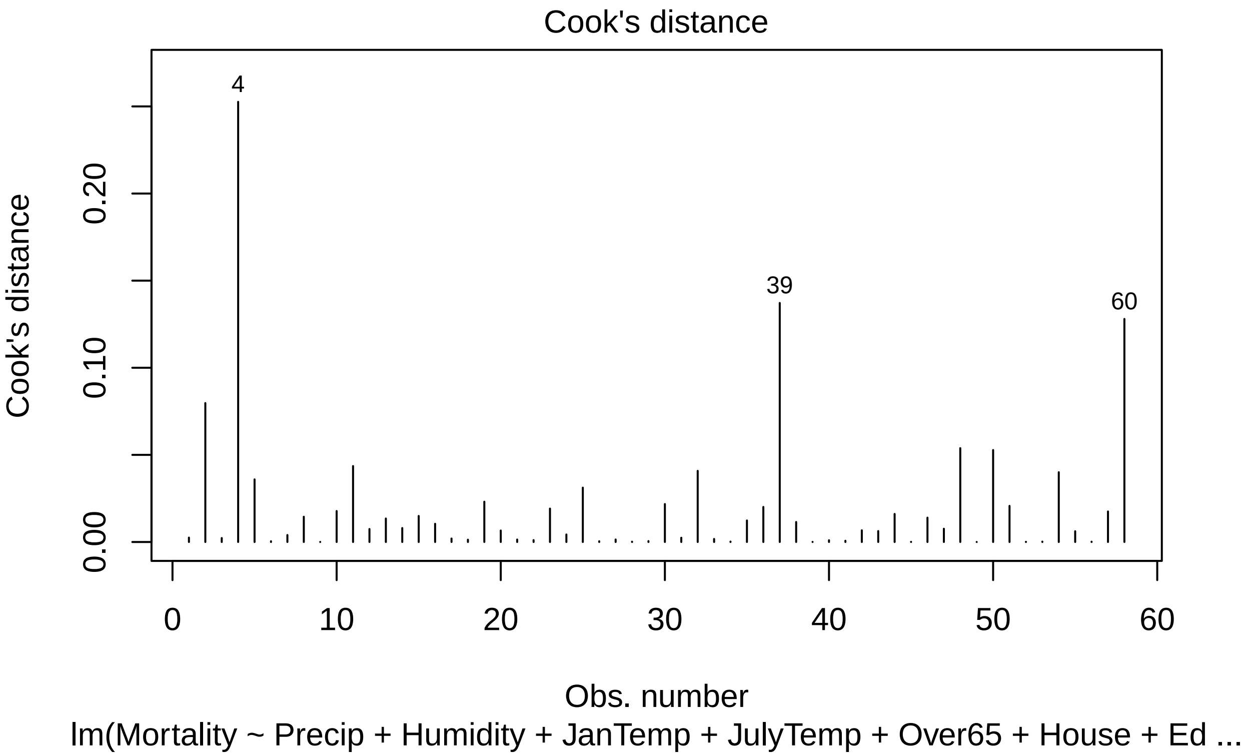


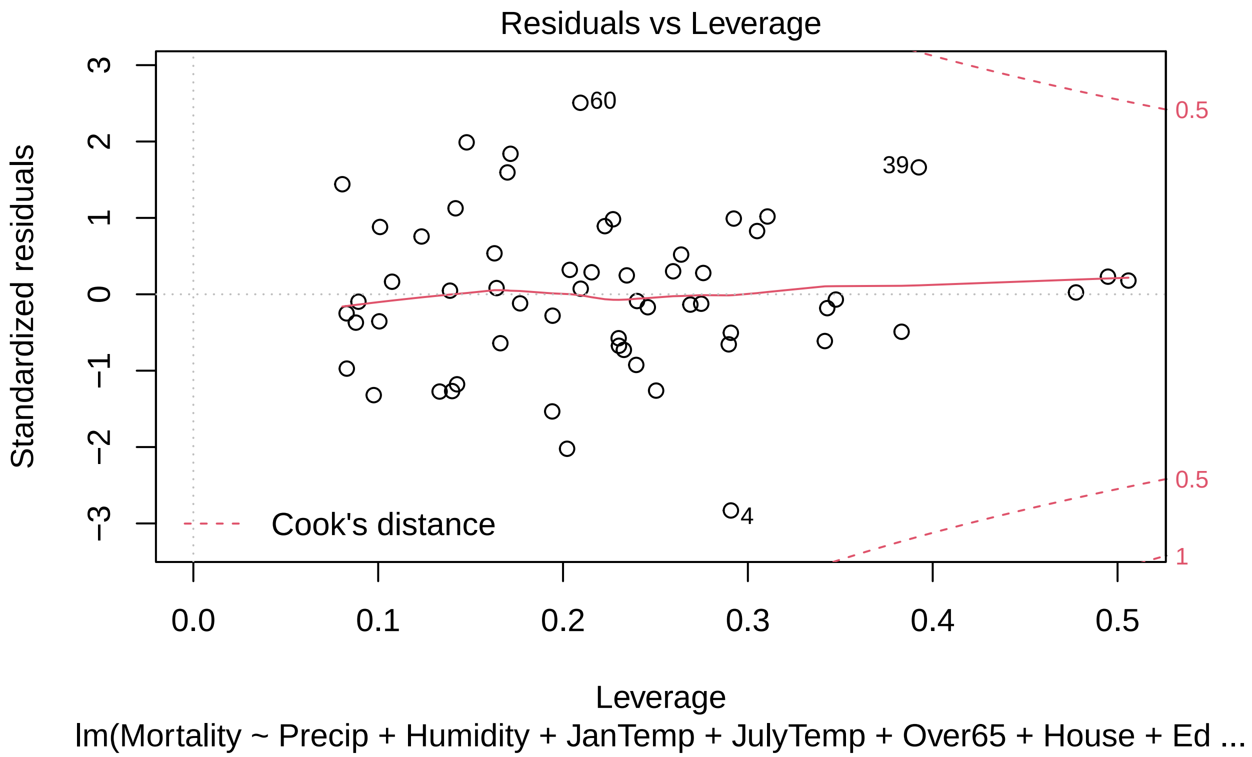
```
ggplot(pm_mutate, aes(Mortality, log(Density))) +
  geom_point() +
  geom_point(data=filter(pm_mutate, case == 20), color="green", size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
# check outliers 3 - weather and demographics
plot(weather_demo_log_lm_no_7_20, which =c(4,5))
```





```
# refit model without case 7, case 20, and case 4
weather_demo_log_lm_no_7_20_4 <- lm(Mortality~Precip+Humidity+JanTemp+JulyTemp+Over65+House+Educ+Sound+l
# no much change, case 4 is not an influential outlier
summary(weather_demo_log_lm_no_7_20_4)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Humidity + JanTemp + JulyTemp +
##     Over65 + House + Educ + Sound + log(Density) + log(NonWhite) +
##     WhiteCol + log(Poor), data = pm, subset = -c(7, 20, 4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.885 -15.652   0.117  16.730  67.764
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   956.4720   456.4389   2.096 0.041913 *
## Precip          2.8963     0.6985   4.147 0.000151 ***
## Humidity       -0.2856     0.9650  -0.296 0.768646
## JanTemp        -0.9414     0.9126  -1.032 0.307870
## JulyTemp       -2.9485     1.6085  -1.833 0.073566 .
## Over65         -5.5723     7.4836  -0.745 0.460473
## House         -37.2492    63.4229  -0.587 0.559994
## Educ          -32.2654    10.1269  -3.186 0.002652 **
## Sound           0.4332     1.7920   0.242 0.810105
## log(Density)   49.8853    14.8231   3.365 0.001594 **
## log(NonWhite)  27.2263     8.6897   3.133 0.003074 **
## WhiteCol        1.6391     1.4178   1.156 0.253891
## log(Poor)      36.7261    47.1617   0.779 0.440309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.98 on 44 degrees of freedom
## Multiple R-squared:  0.8001, Adjusted R-squared:  0.7456
## F-statistic: 14.67 on 12 and 44 DF,  p-value: 1.163e-11
```

```
# check collinearity 1 - weather and demographics
vif(weather_demo_log_lm_no_7_20)
```

```
##        Precip      Humidity       JanTemp      JulyTemp        Over65
##      2.689407      1.594603      3.871611      3.446249      7.117445
##         House          Educ         Sound  log(Density) log(NonWhite)
##      3.964616      3.646570      4.777184      1.524412      3.442752
##      WhiteCol     log(Poor)
##      2.488016      7.477916
```

```
# anova 1 - weather and demographics
small_lm1 <- lm(Mortality~Precip+Educ+log(Density)+log(NonWhite)+log(Poor), data=pm, subset=-c(7, 20))
big_lm1 <- lm(Mortality~Precip+Humidity+JanTemp+JulyTemp+Over65+House+Educ+Sound+log(Density)+log(NonWhi
# no term is significant
anova(small_lm1, big_lm1)
```

```
## Analysis of Variance Table
##
## Model 1: Mortality ~ Precip + Educ + log(Density) + log(NonWhite) + log(Poor)
## Model 2: Mortality ~ Precip + Humidity + JanTemp + JulyTemp + Over65 +
##     House + Educ + Sound + log(Density) + log(NonWhite) + WhiteCol +
##     log(Poor)
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     52 59204
## 2     45 51373  7    7831.2 0.98 0.4576
```

```
# log(Poor) is not significant
summary(small_lm1)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(Density) + log(NonWhite) +
##     log(Poor), data = pm, subset = -c(7, 20))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.543  -17.239   -1.446   16.873   73.173
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   450.1596   201.2978   2.236   0.0296 *
## Precip          2.4450     0.5788   4.225 9.67e-05 ***
## Educ          -18.5491     7.3283  -2.531   0.0144 *
## log(Density)   62.6718    14.3338   4.372 5.92e-05 ***
## log(NonWhite)  27.5157     6.4002   4.299 7.55e-05 ***
## log(Poor)      12.9288    25.7944   0.501   0.6183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 33.74 on 52 degrees of freedom
## Multiple R-squared:  0.7321, Adjusted R-squared:  0.7064
## F-statistic: 28.43 on 5 and 52 DF,  p-value: 9.127e-14
```

```r
# anova 2 - weather and demographics
small_lm2 <- lm(Mortality~Precip+Educ+log(Density)+log(NonWhite), data=pm, subset=-c(7, 20))
big_lm2 <- lm(Mortality~Precip+Humidity+JanTemp+JulyTemp+Over65+House+Educ+Sound+log(Density)+log(NonWhi
# no term is significant
anova(small_lm2, big_lm2)
```

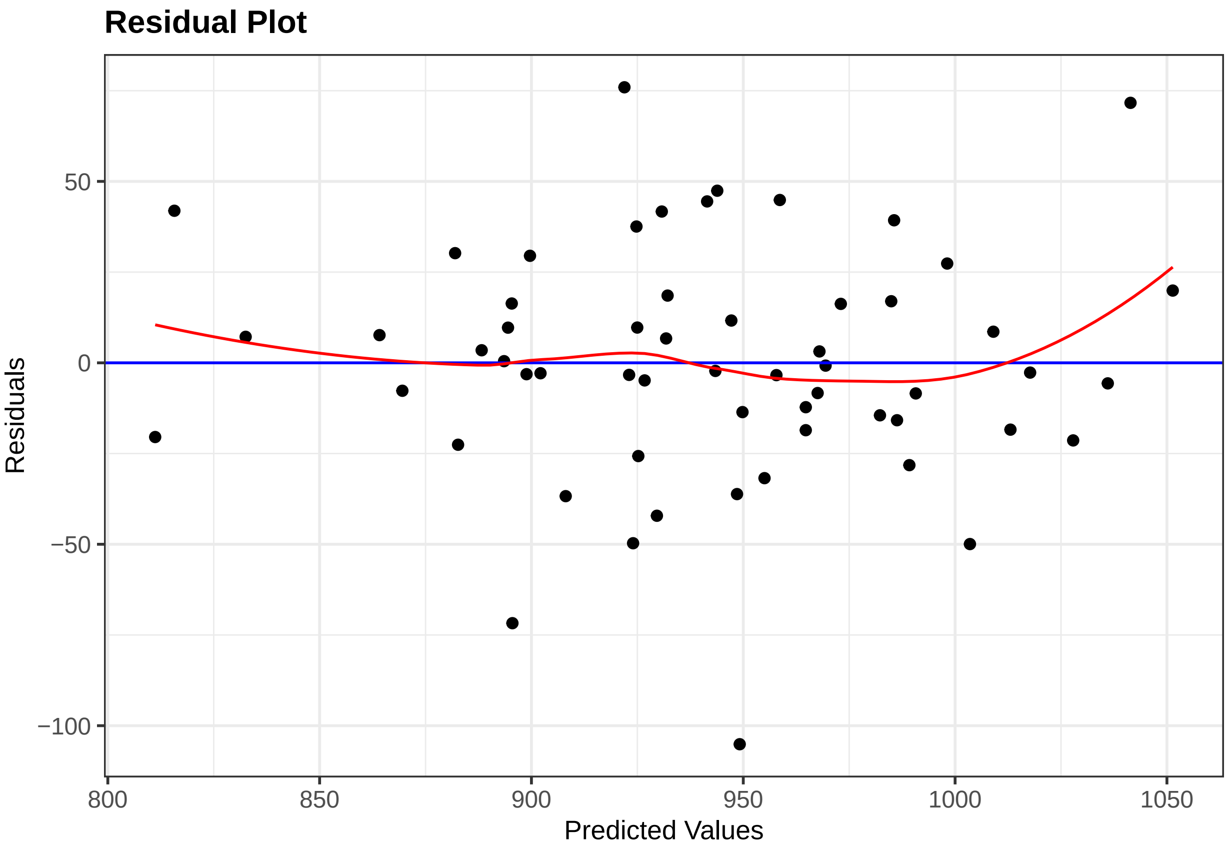
```
## Analysis of Variance Table
##
## Model 1: Mortality ~ Precip + Educ + log(Density) + log(NonWhite)
## Model 2: Mortality ~ Precip + Humidity + JanTemp + JulyTemp + Over65 +
##     House + Educ + Sound + log(Density) + log(NonWhite) + WhiteCol +
##     log(Poor)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     53 59490
## 2     45 51373  8    8117.3 0.8888 0.5336
```

```r
# every term is significant
summary(small_lm2)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(Density) + log(NonWhite),
##     data = pm, subset = -c(7, 20))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -105.108  -17.766   -2.466   16.816   75.935
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   517.0763   149.5869   3.457  0.00109 **
## Precip          2.4753     0.5715   4.331 6.63e-05 ***
## Educ          -20.1045     6.5918  -3.050  0.00357 **
## log(Density)   60.1122    13.2984   4.520 3.50e-05 ***
## log(NonWhite)  29.3164     5.2593   5.574 8.54e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.5 on 53 degrees of freedom
## Multiple R-squared:  0.7309, Adjusted R-squared:  0.7105
## F-statistic: 35.98 on 4 and 53 DF,  p-value: 1.598e-14
```

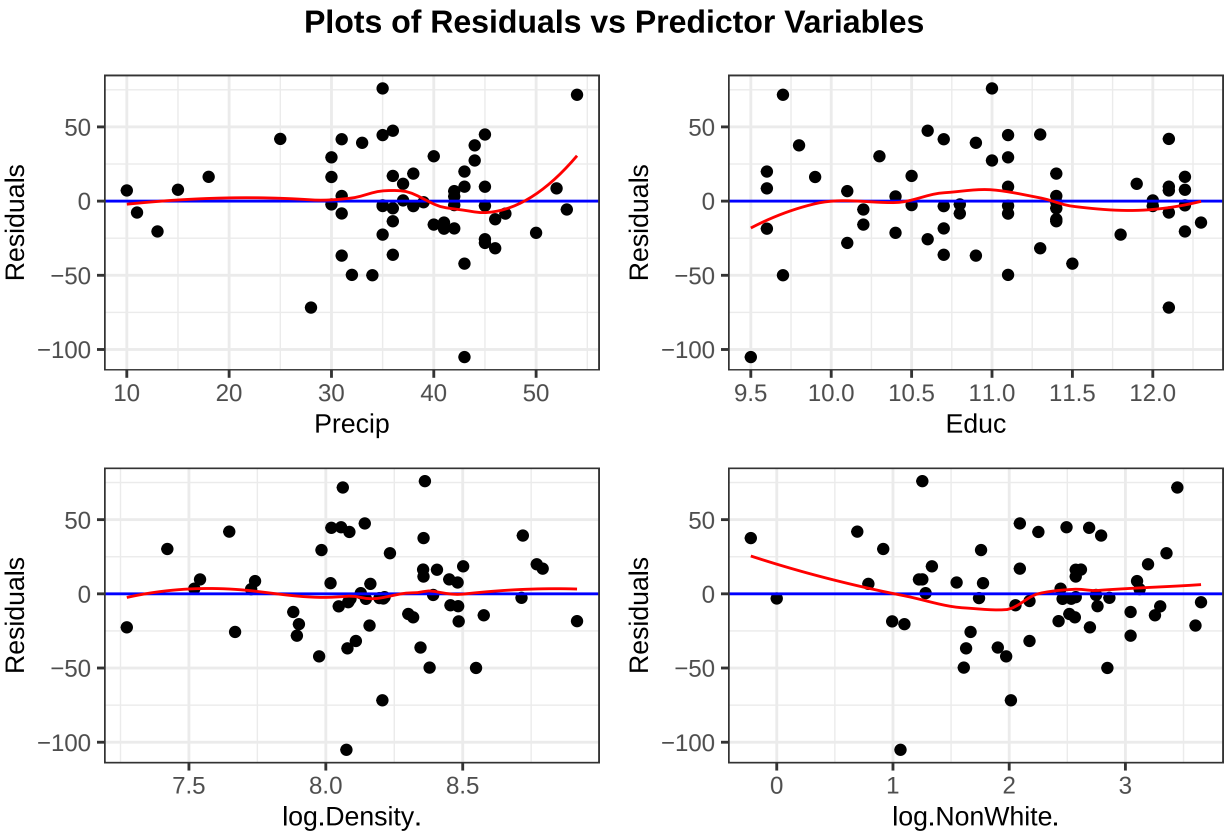
```r
# check assumptions 4 - weather and demographics
# residuals plot
resid_panel(small_lm2, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

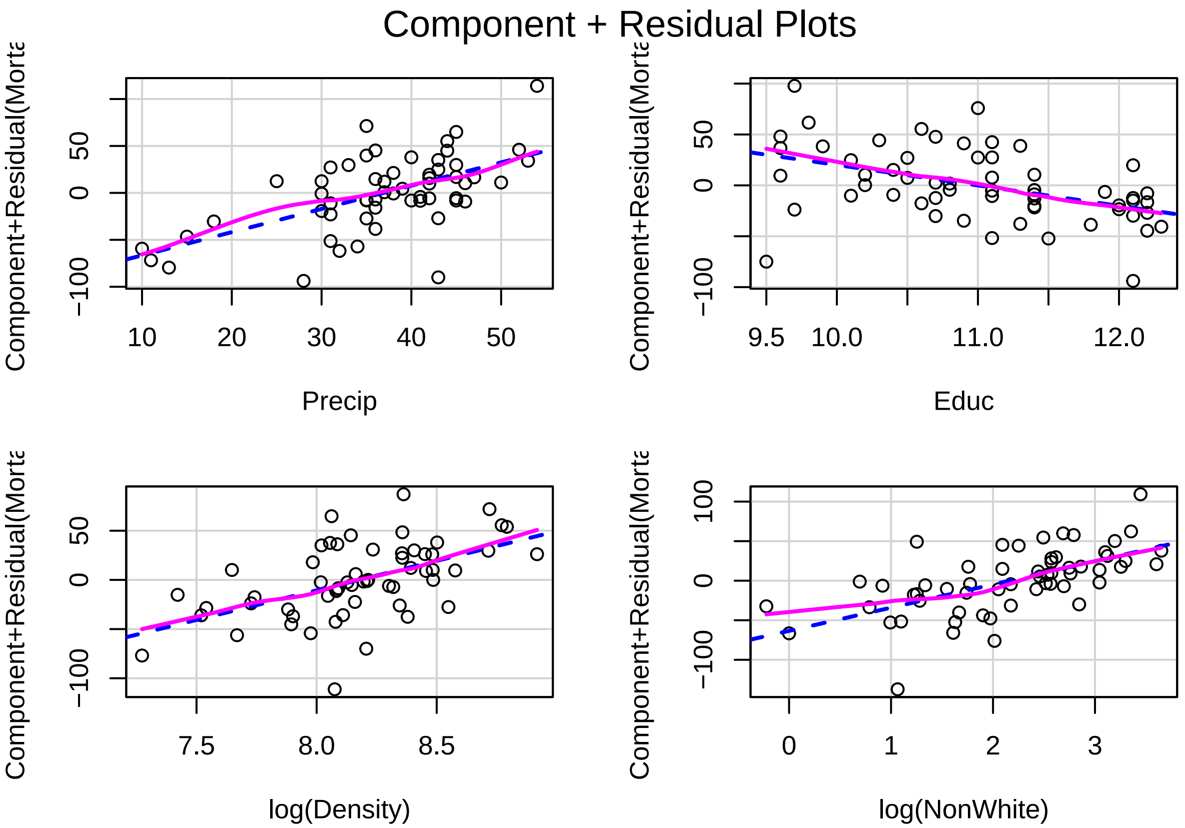


```
# residuals of each predictor
resid_xpanel(small_lm2, smoother = TRUE)
```

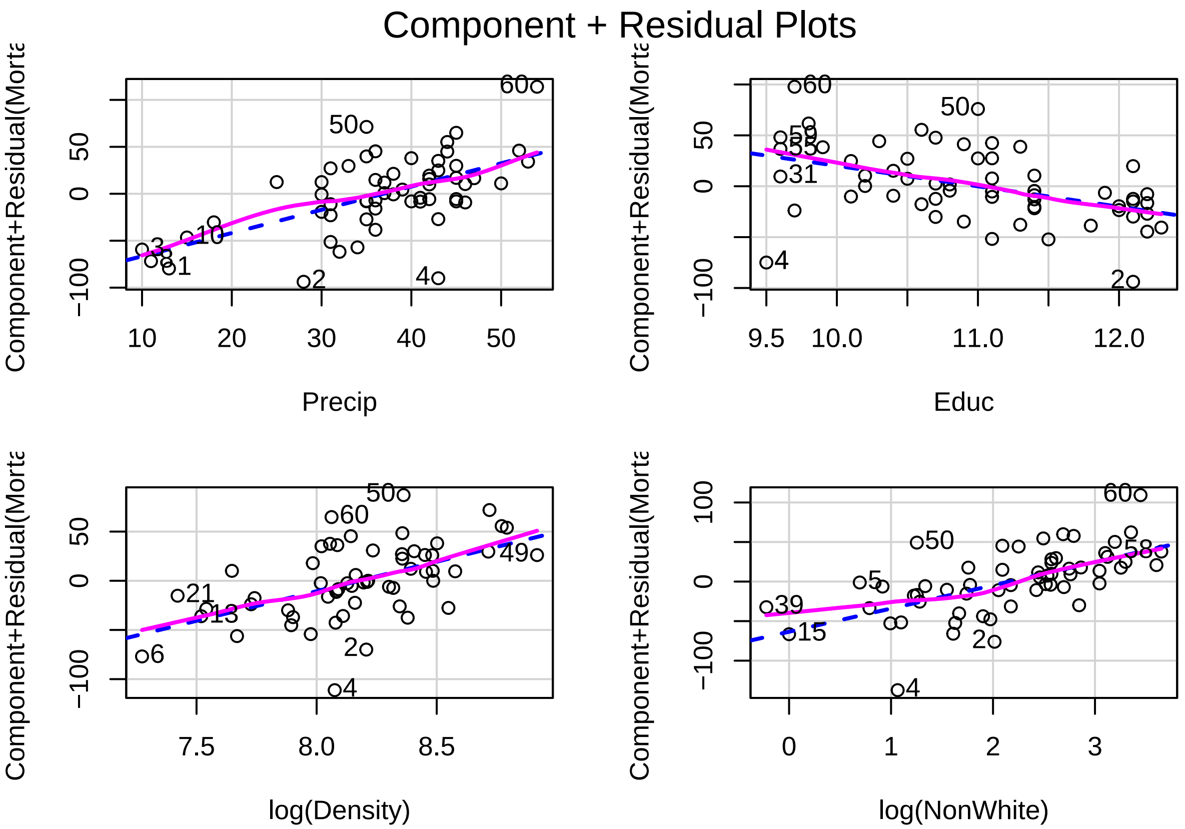
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



```
# partial residuals
crp(small_lm2)
```



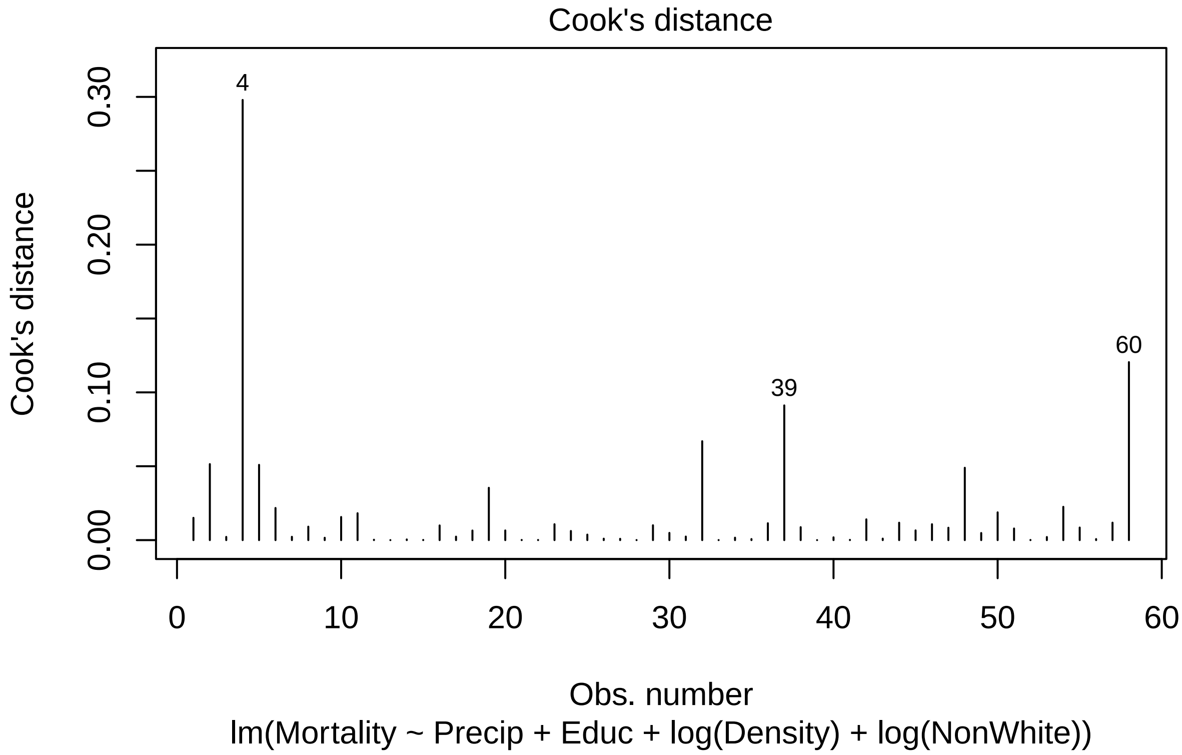
```r
crp(small_lm2, id = list(n = 4))
```

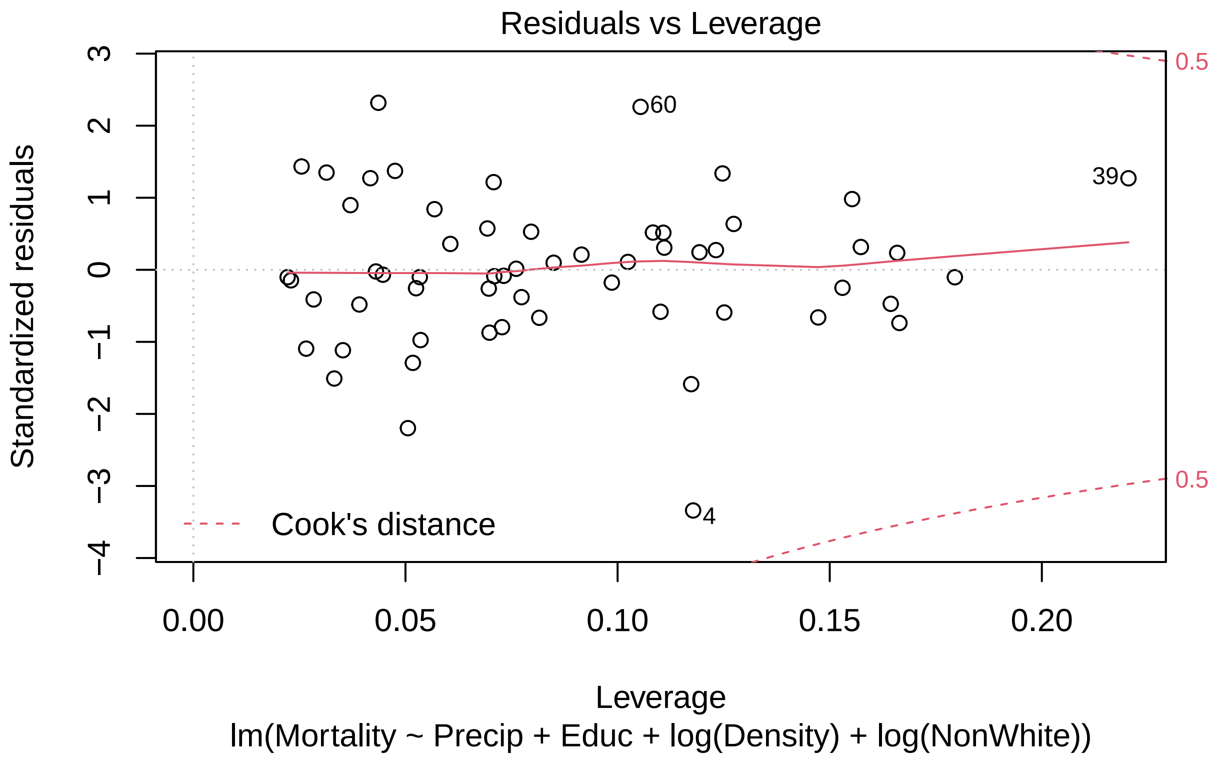


```
# check collinearity 2 - weather and demographics
vif(small_lm2)
```

```
##        Precip          Educ  log(Density) log(NonWhite)
##      1.552503      1.469068      1.051960      1.081408
```

```
# check outliers 4 - weather and demographics
plot(small_lm2, which =c(4,5))
```

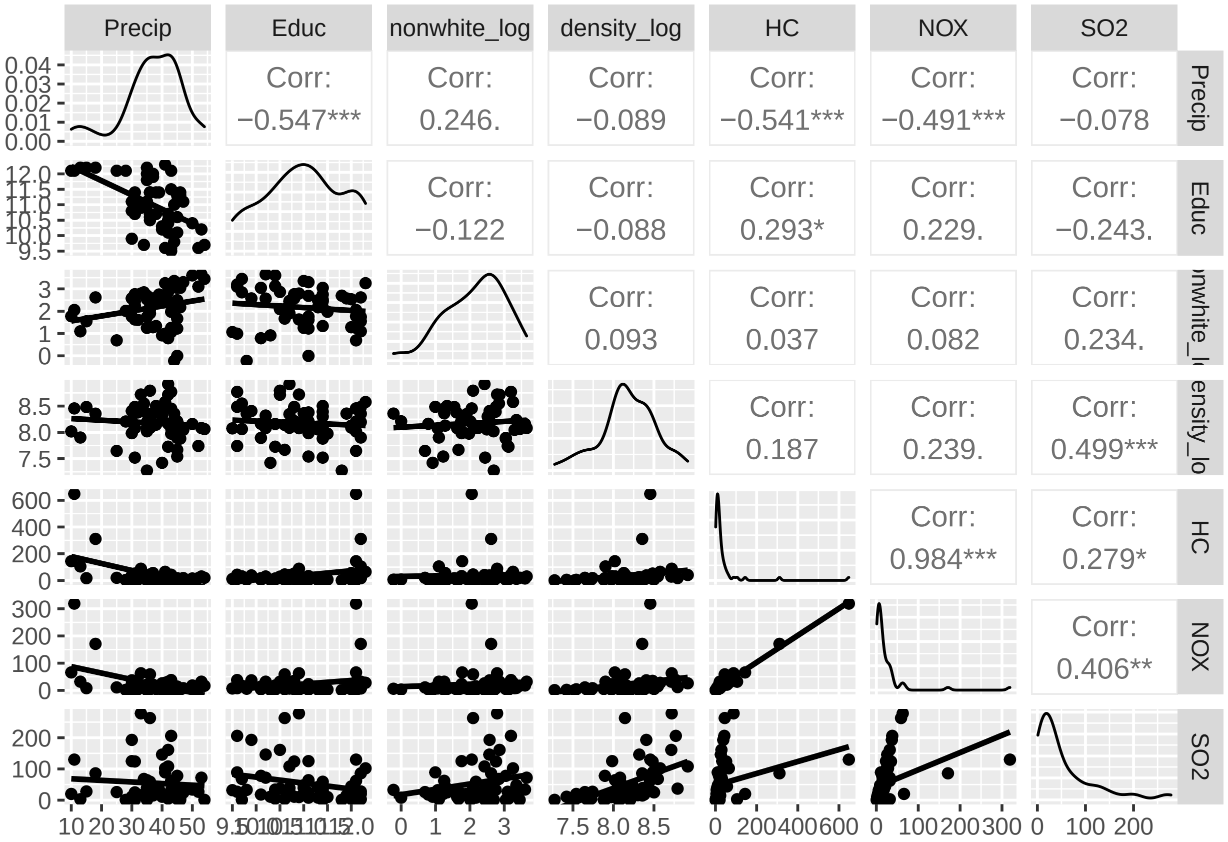




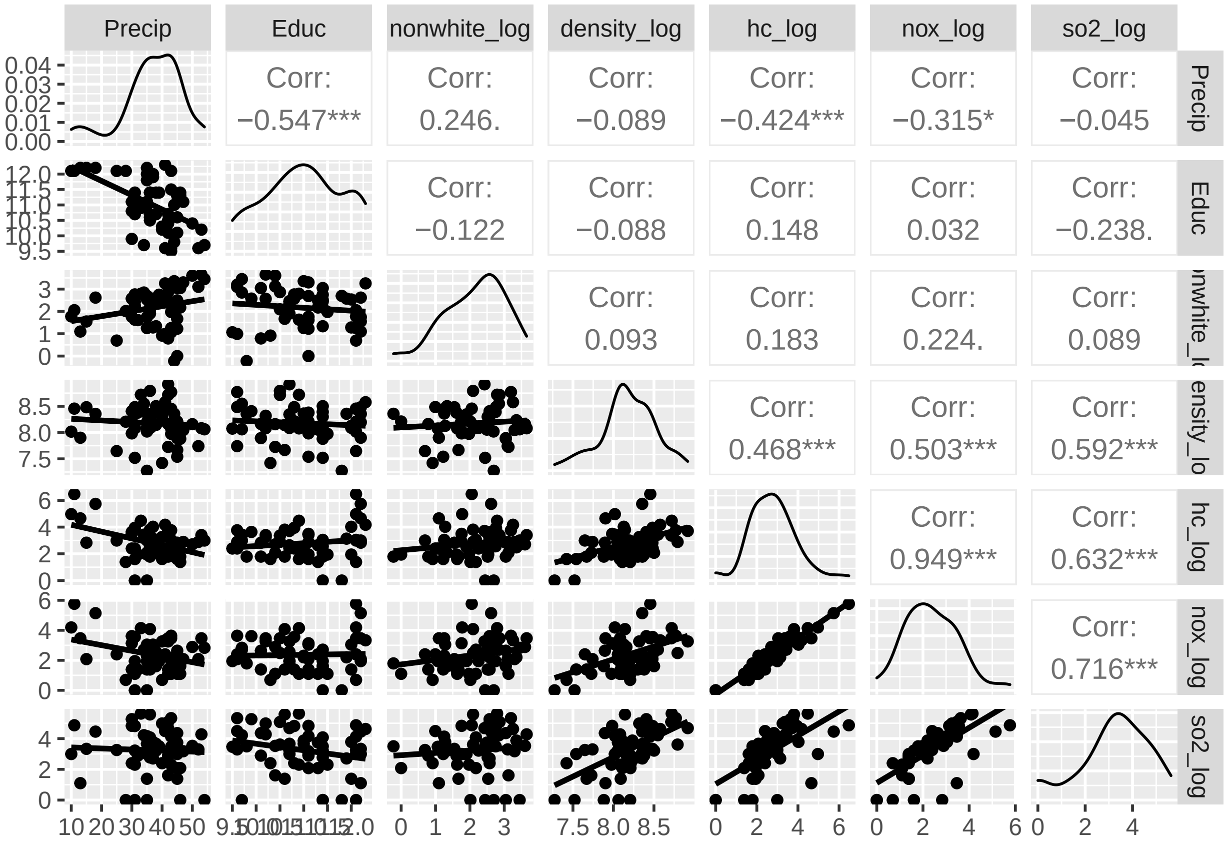
```
# refit model without case 7, case 20, and case 4
small_lm2_no_4 <- lm(Mortality~Precip+Educ+log(Density)+log(NonWhite), data=pm, subset=-c(7, 20, 4))
# no much change, case 4 is not an influential outlier
summary(small_lm2_no_4)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(Density) + log(NonWhite),
##     data = pm, subset = -c(7, 20, 4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.558 -19.270   0.234  16.437  71.137
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    608.5168   136.4138   4.461 4.40e-05 ***
## Precip           2.4226     0.5129   4.724 1.80e-05 ***
## Educ           -25.6798     6.0997  -4.210 0.000101 ***
## log(Density)    57.8130    11.9451   4.840 1.20e-05 ***
## log(NonWhite)   25.9305     4.8046   5.397 1.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.05 on 52 degrees of freedom
## Multiple R-squared:  0.7776, Adjusted R-squared:  0.7605
## F-statistic: 45.47 on 4 and 52 DF,  p-value: 2.238e-16
```

```
# Adding pollution variables
# ggpairs - base + pollutions
without_outlier_pm <- pm_mutate %>% filter(CITY != "Miami, FL" & CITY != "York, PA")
without_outlier_pm %>%
  mutate(
      density_log = log(Density),
      nonwhite_log = log(NonWhite)
    ) %>%
  ggpairs(columns = c("Precip", "Educ", "nonwhite_log", "density_log", "HC","NOX", "SO2"),
          lower = list(continuous =  wrap("smooth", se = FALSE)))
```



```
# ggpairs - base + log(pollution variables)
without_outlier_pm %>%
  mutate(
      hc_log = log(HC),
      nox_log = log(NOX),
      so2_log = log(SO2),
      density_log = log(Density),
      nonwhite_log = log(NonWhite)
    ) %>%
  ggpairs(columns = c("Precip", "Educ", "nonwhite_log", "density_log", "hc_log","nox_log", "so2_log"),
          lower = list(continuous =  wrap("smooth", se = FALSE)))
```

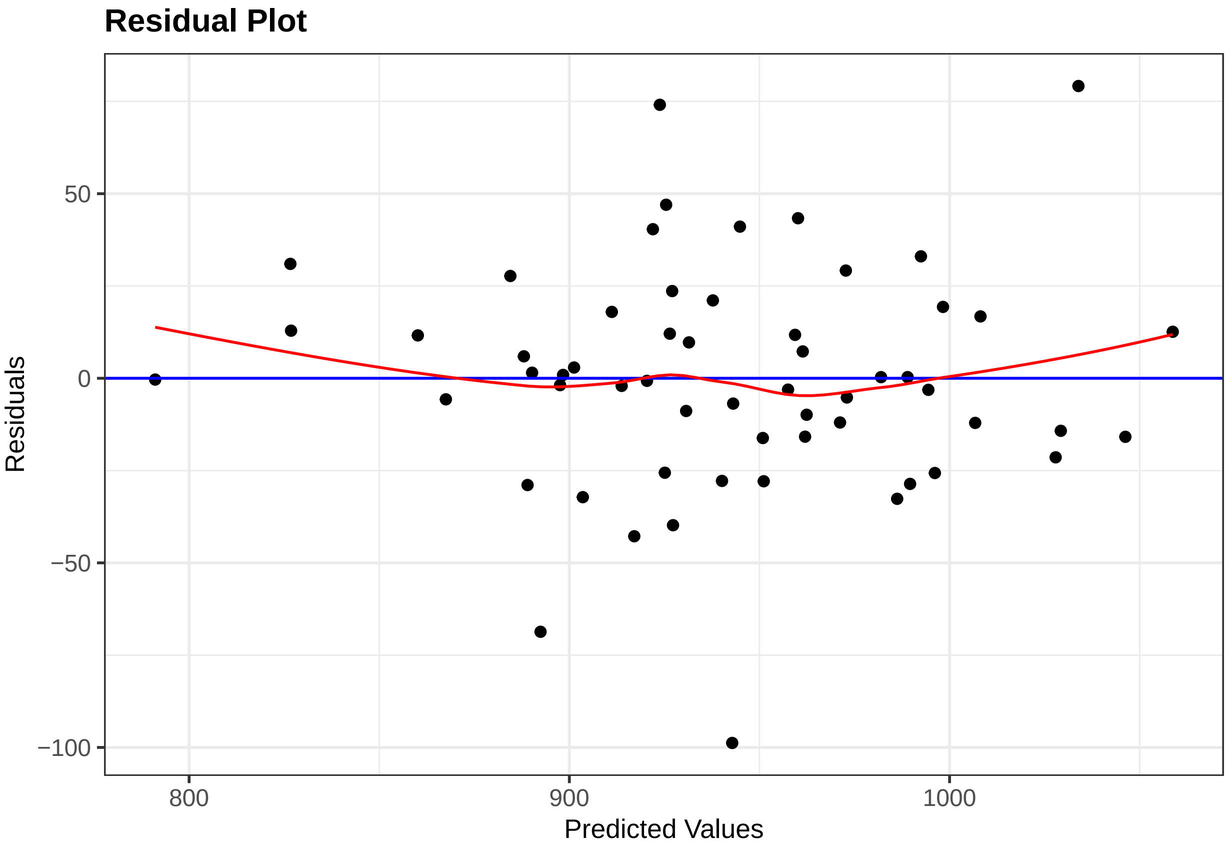


```r
# untransformed model - pollution
pollution_unlog_lm <- lm(Mortality~Precip+Educ+log(NonWhite)+log(Density)+HC+NOX+SO2, data=pm, subset=-
summary(pollution_unlog_lm)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(NonWhite) + log(Density) +
##     HC + NOX + SO2, data = pm, subset = -c(7, 20))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.783 -15.844  -0.519  15.792  79.169
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   569.90981  146.28340   3.896 0.000291 ***
## Precip          2.50230    0.66217   3.779 0.000420 ***
## Educ          -15.49872    6.74896  -2.296 0.025876 *
## log(NonWhite)  26.26542    5.30628   4.950 8.84e-06 ***
## log(Density)   47.15591   14.51994   3.248 0.002081 **
## HC             -0.59082    0.41219  -1.433 0.157971
## NOX             1.19070    0.83723   1.422 0.161183
## SO2             0.06139    0.12352   0.497 0.621365
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.34 on 50 degrees of freedom
## Multiple R-squared:  0.7634, Adjusted R-squared:  0.7303
## F-statistic: 23.05 on 7 and 50 DF,  p-value: 1.302e-13
```

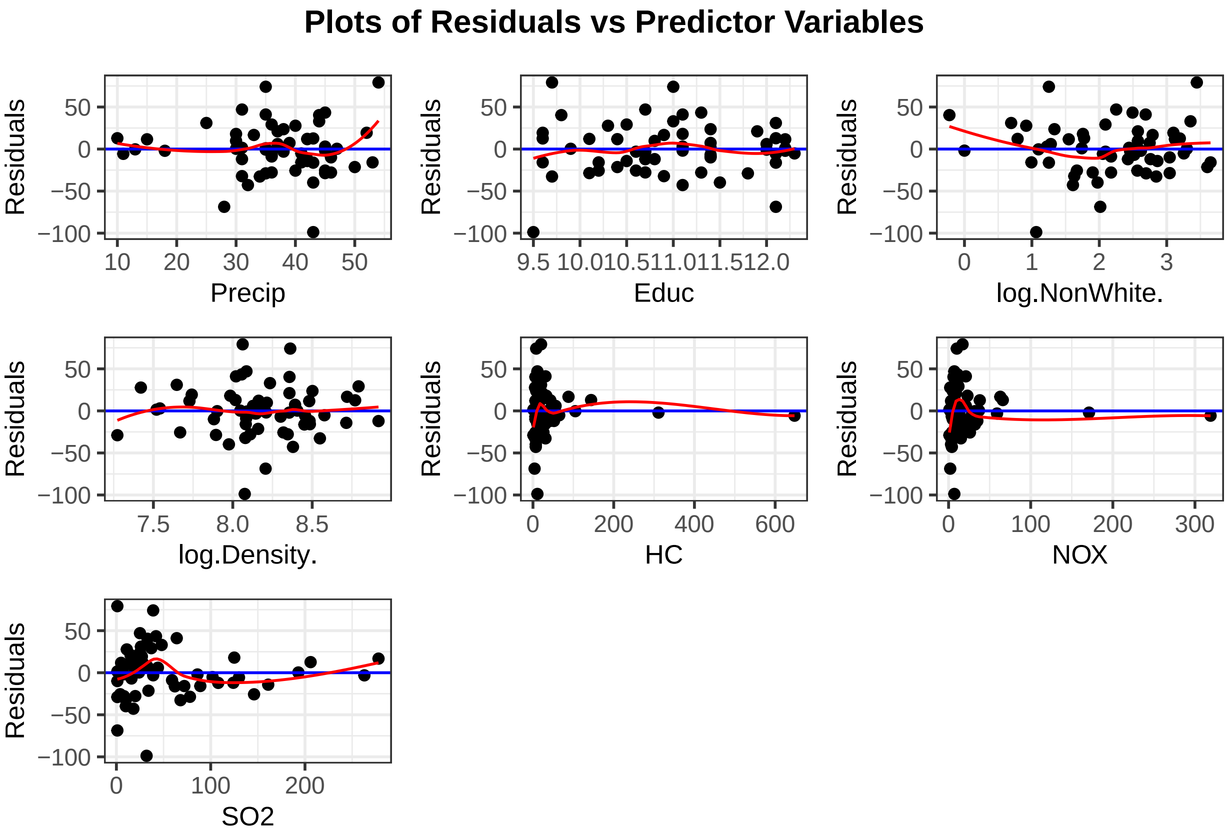
```
# check assumptions 1 - pollution
# residuals plot
resid_panel(pollution_unlog_lm, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

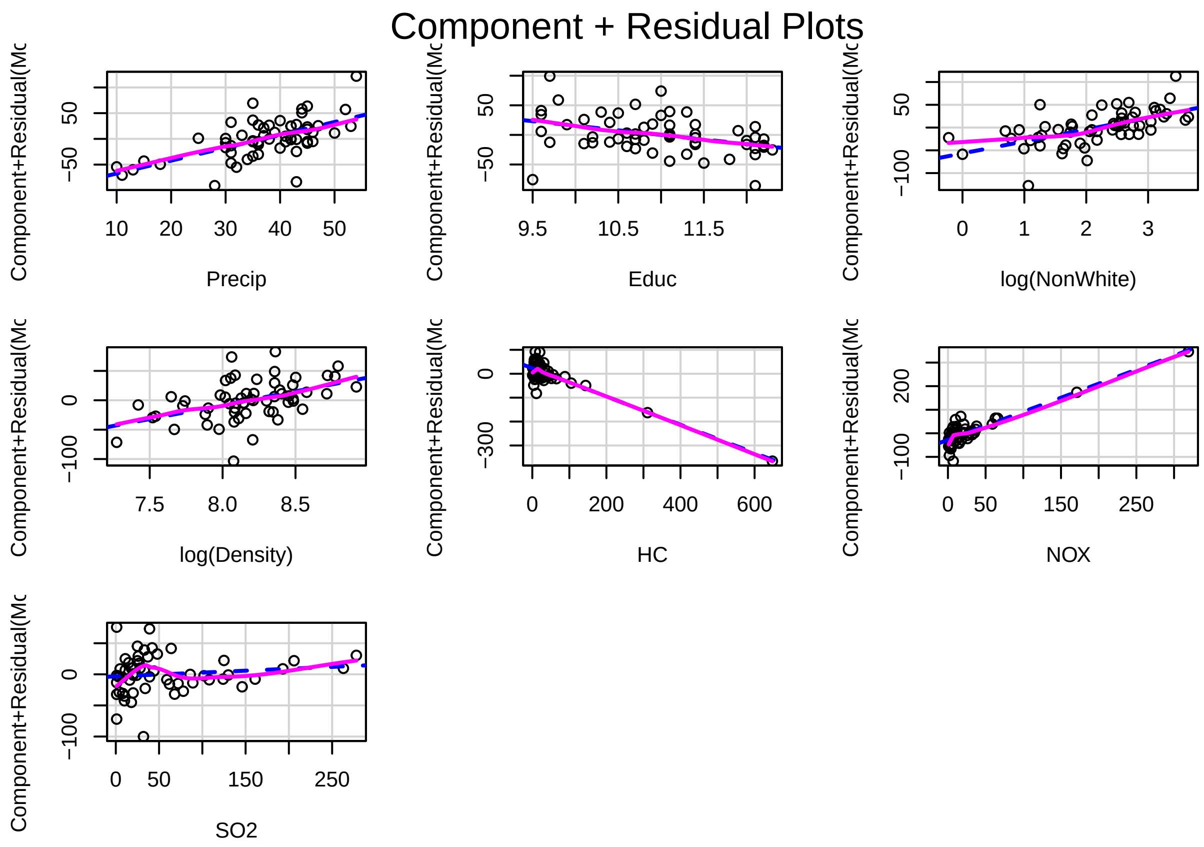


```
# residuals of each predictor
resid_xpanel(pollution_unlog_lm, smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



```
# partial residuals
crp(pollution_unlog_lm)
```



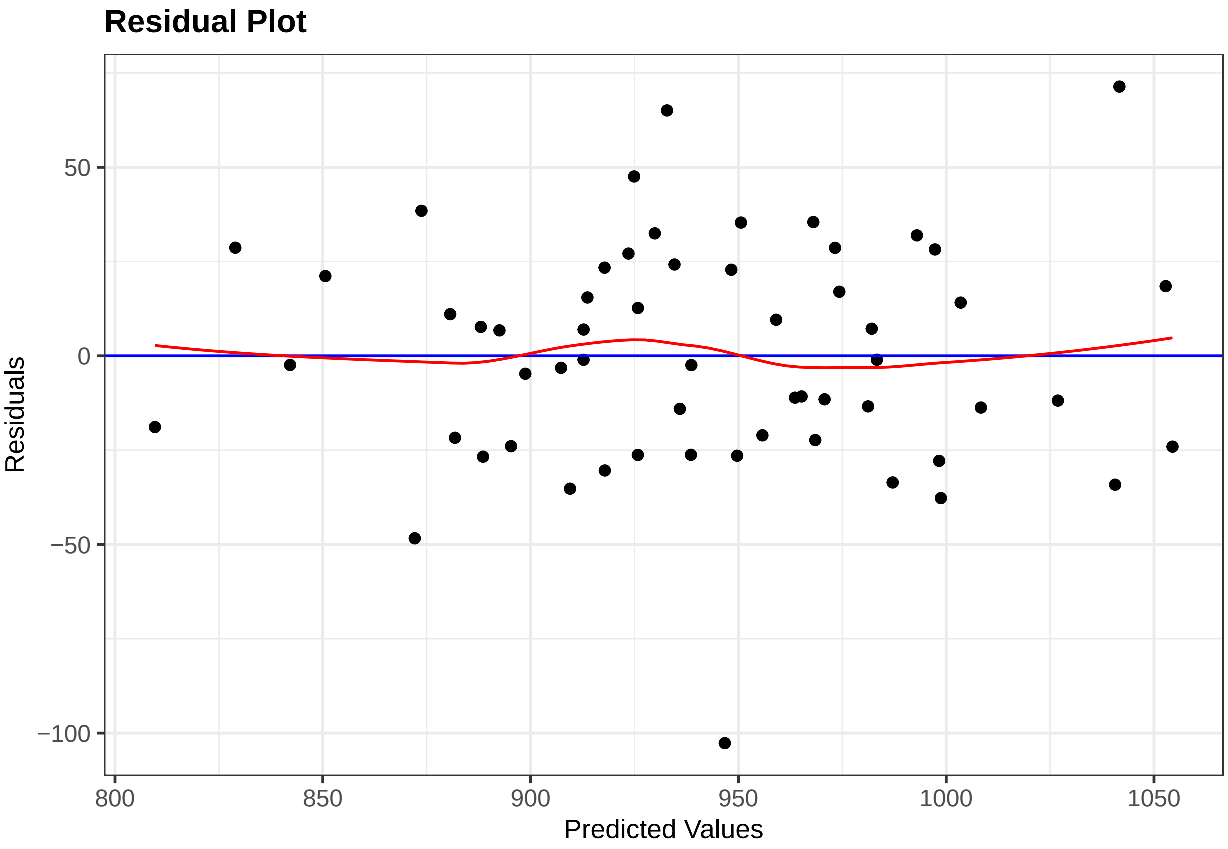
```
# transformed model - pollution
pollution_log_lm <- lm(Mortality~Precip+Educ+log(NonWhite)+log(Density)+log(HC)+log(NOX)+log(SO2), data=
# log(HC) and log(NOX) are not significant
summary(pollution_log_lm)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(NonWhite) + log(Density) +
##     log(HC) + log(NOX) + log(SO2), data = pm, subset = -c(7,
##     20))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -102.67  -22.17   -1.73   22.41   71.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   601.4947   147.6629   4.073 0.000165 ***
## Precip          2.7871     0.6188   4.504 4.02e-05 ***
## Educ          -15.8655     6.6278  -2.394 0.020469 *
## log(NonWhite)  25.8518     5.3983   4.789 1.53e-05 ***
## log(Density)   41.2913    15.5150   2.661 0.010436 *
## log(HC)       -17.8446    12.8026  -1.394 0.169537
## log(NOX)       26.4841    13.4747   1.965 0.054926 .
## log(SO2)        1.8459     4.9381   0.374 0.710127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 31.92 on 50 degrees of freedom
## Multiple R-squared:  0.7695, Adjusted R-squared:  0.7372
## F-statistic: 23.85 on 7 and 50 DF,  p-value: 6.93e-14
```

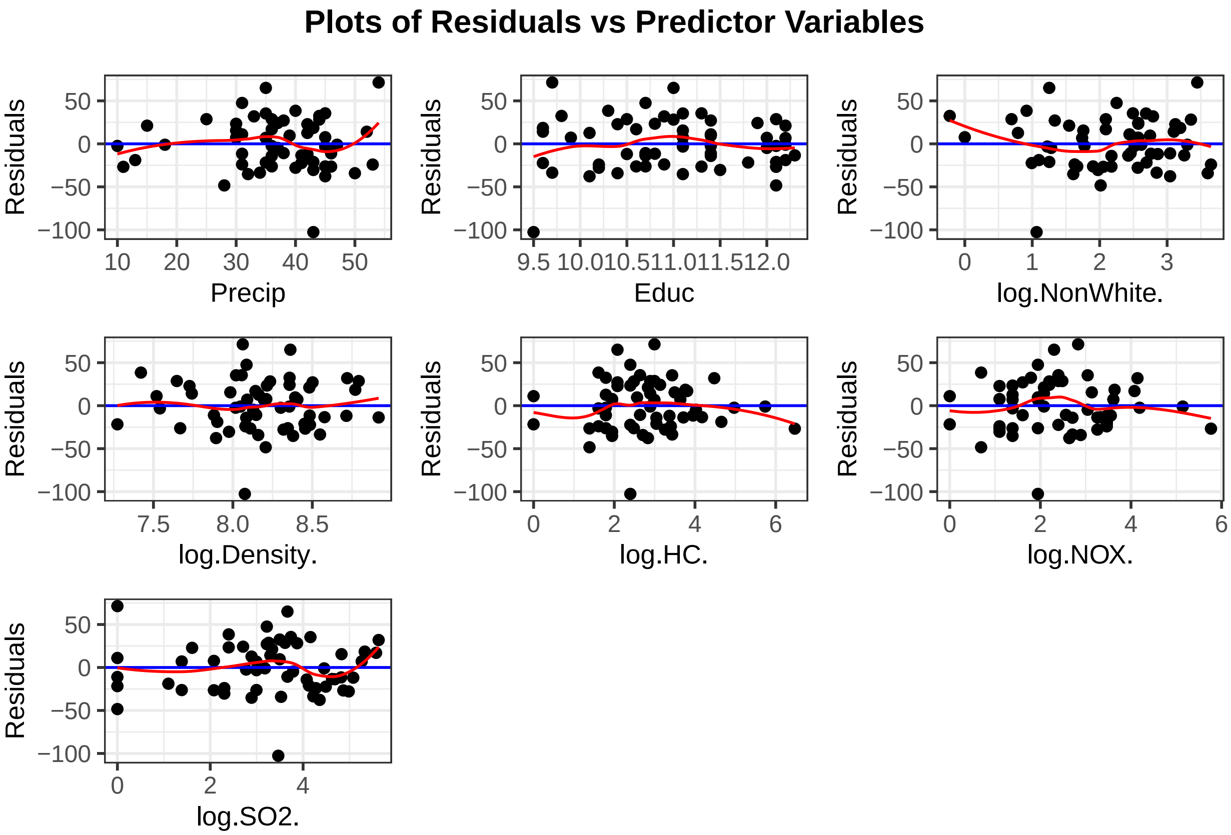
```
# check assumptions 2 - pollution
# residuals plot
resid_panel(pollution_log_lm, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

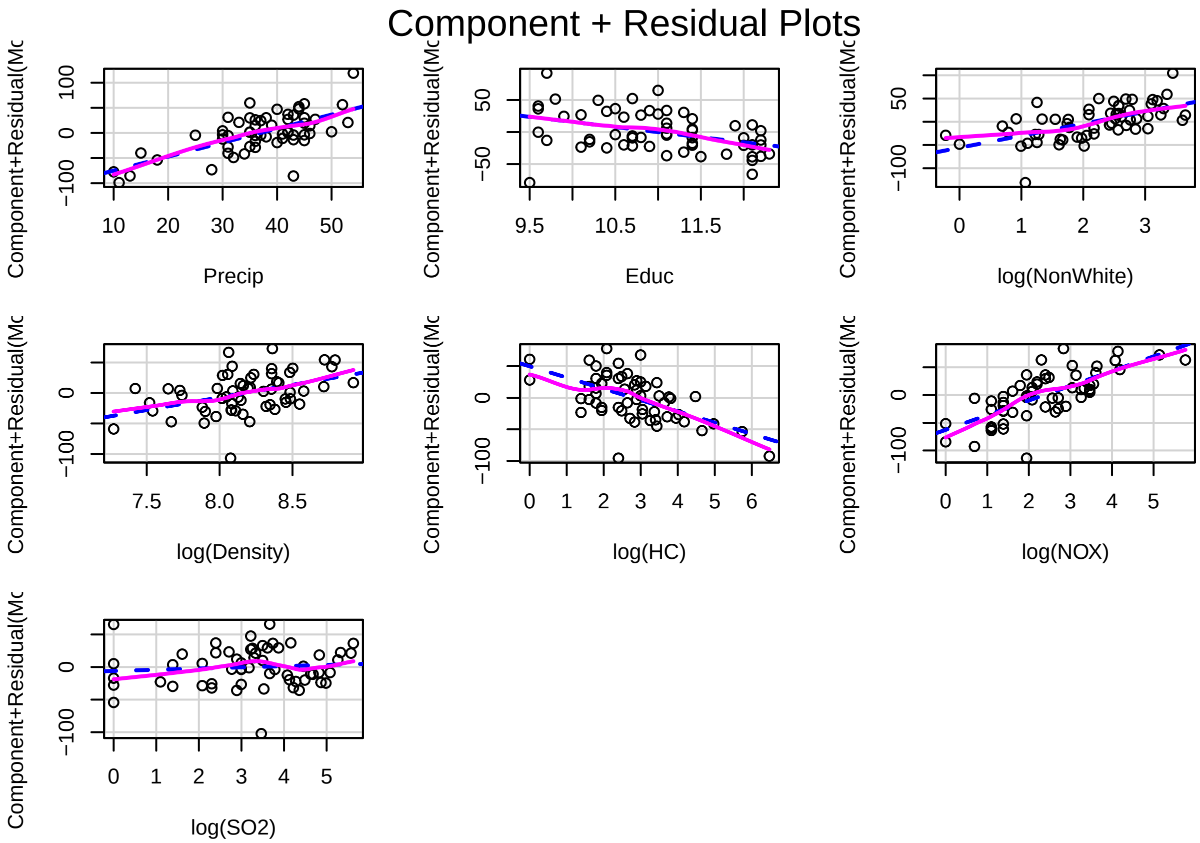


```
# residuals of each predictor
resid_xpanel(pollution_log_lm, smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



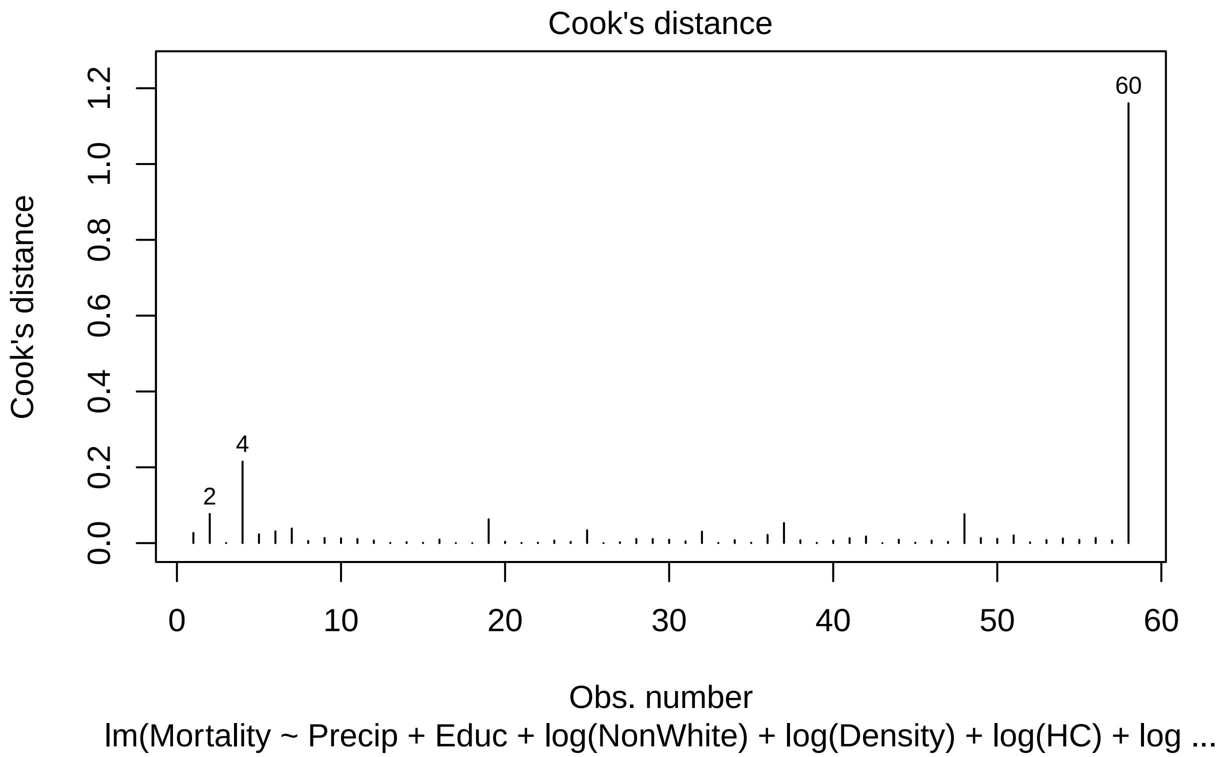
```
# partial residuals
crp(pollution_log_lm)
```

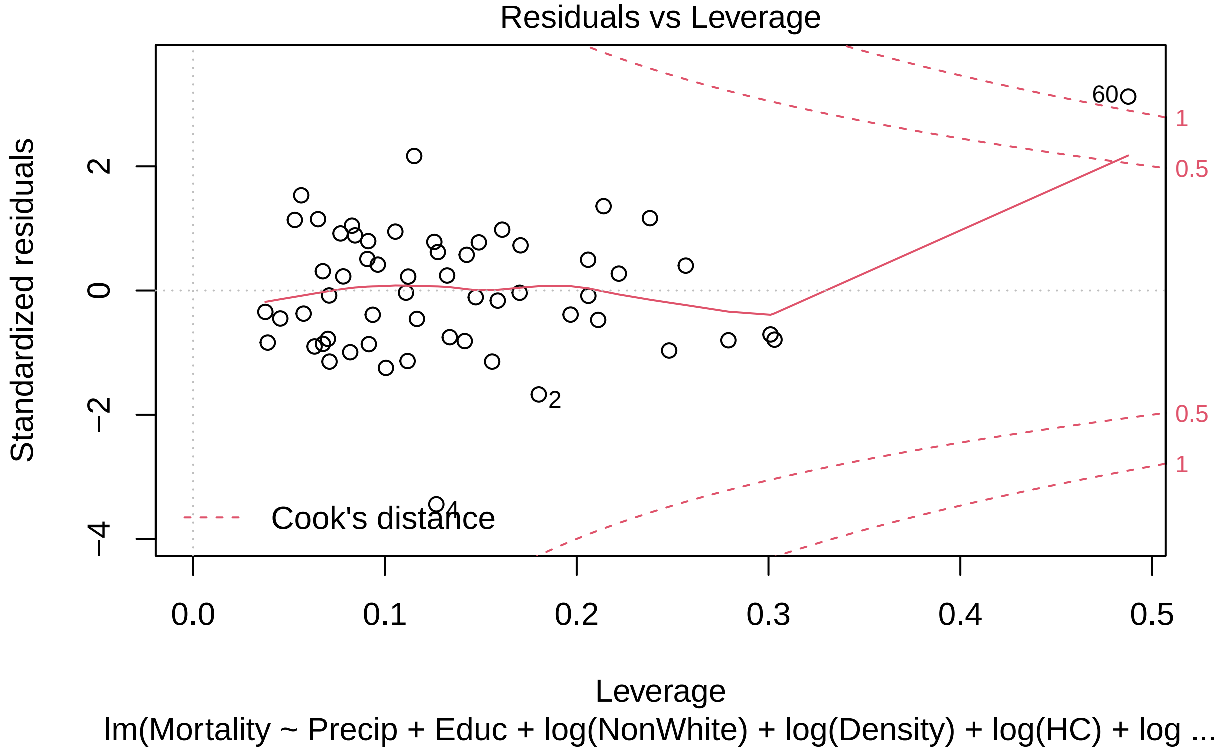


```
# check collinearity 1 - pollution
vif(pollution_log_lm)
```

```
##        Precip          Educ log(NonWhite)  log(Density)       log(HC)
##      2.005035      1.636111      1.255122      1.577392     12.594008
##      log(NOX)      log(SO2)
##     13.755729      2.908046
```

```
# check outliers 1 - pollution
plot(pollution_log_lm, which =c(4,5))
```





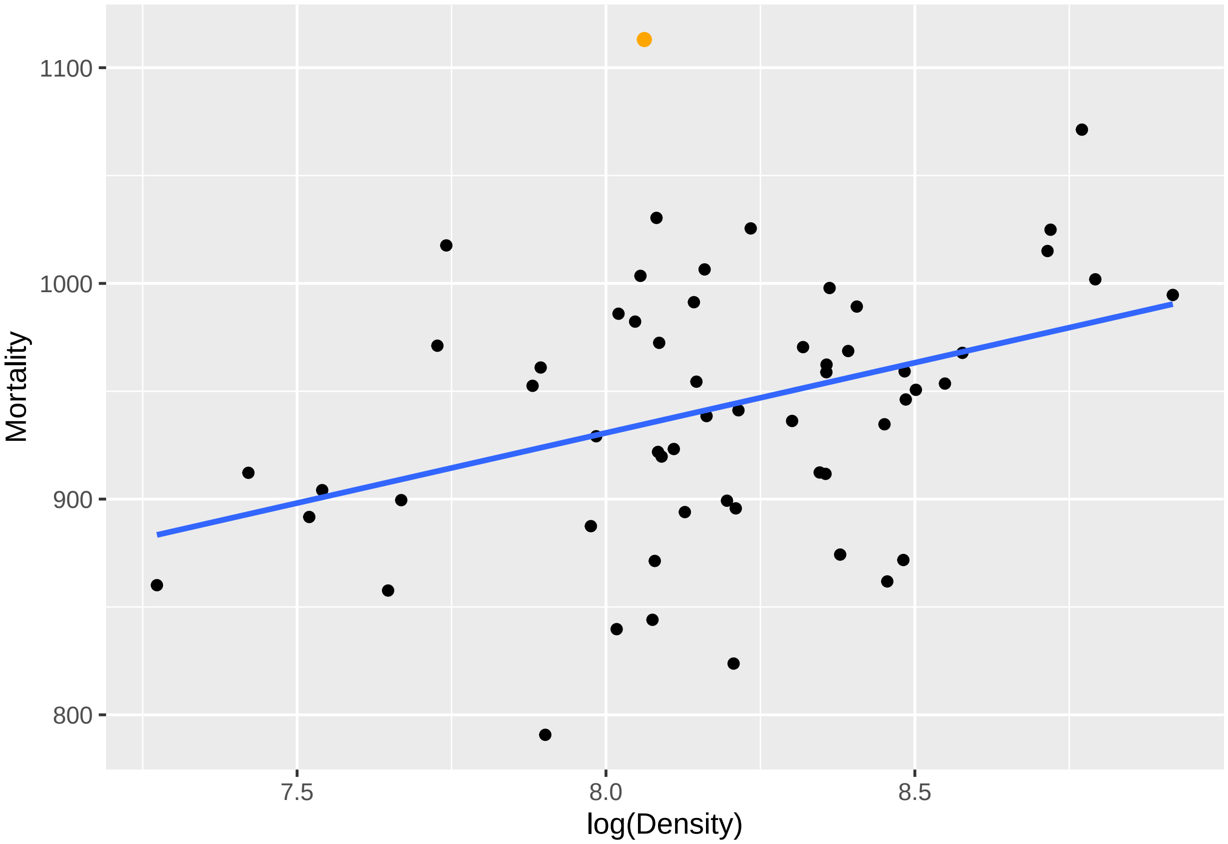
```
# refit model without case 7, case 20, and case 60
pollution_log_lm_no_7_20_60 <- lm(Mortality~Precip+Educ+log(NonWhite)+log(Density)+log(HC)+log(NOX)+log
# Educ is not significant anymore, case 60 is influential
summary(pollution_log_lm_no_7_20_60)
```

```
##
```

```
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(NonWhite) + log(Density) +
##     log(HC) + log(NOX) + log(SO2), data = pm, subset = -c(7,
##     20, 60))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.136 -20.481   1.162  21.447  62.192
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    598.821    133.819   4.475 4.56e-05 ***
## Precip           2.225      0.584   3.811 0.000387 ***
## Educ            -8.885      6.338  -1.402 0.167273
## log(NonWhite)   26.349      4.894   5.384 2.05e-06 ***
## log(Density)    32.923     14.268   2.307 0.025298 *
## log(HC)        -17.561     11.602  -1.514 0.136553
## log(NOX)        13.176     12.807   1.029 0.308617
## log(SO2)        14.902      5.863   2.542 0.014243 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.93 on 49 degrees of freedom
## Multiple R-squared:  0.7857, Adjusted R-squared:  0.7551
## F-statistic: 25.66 on 7 and 49 DF,  p-value: 2.432e-14
```

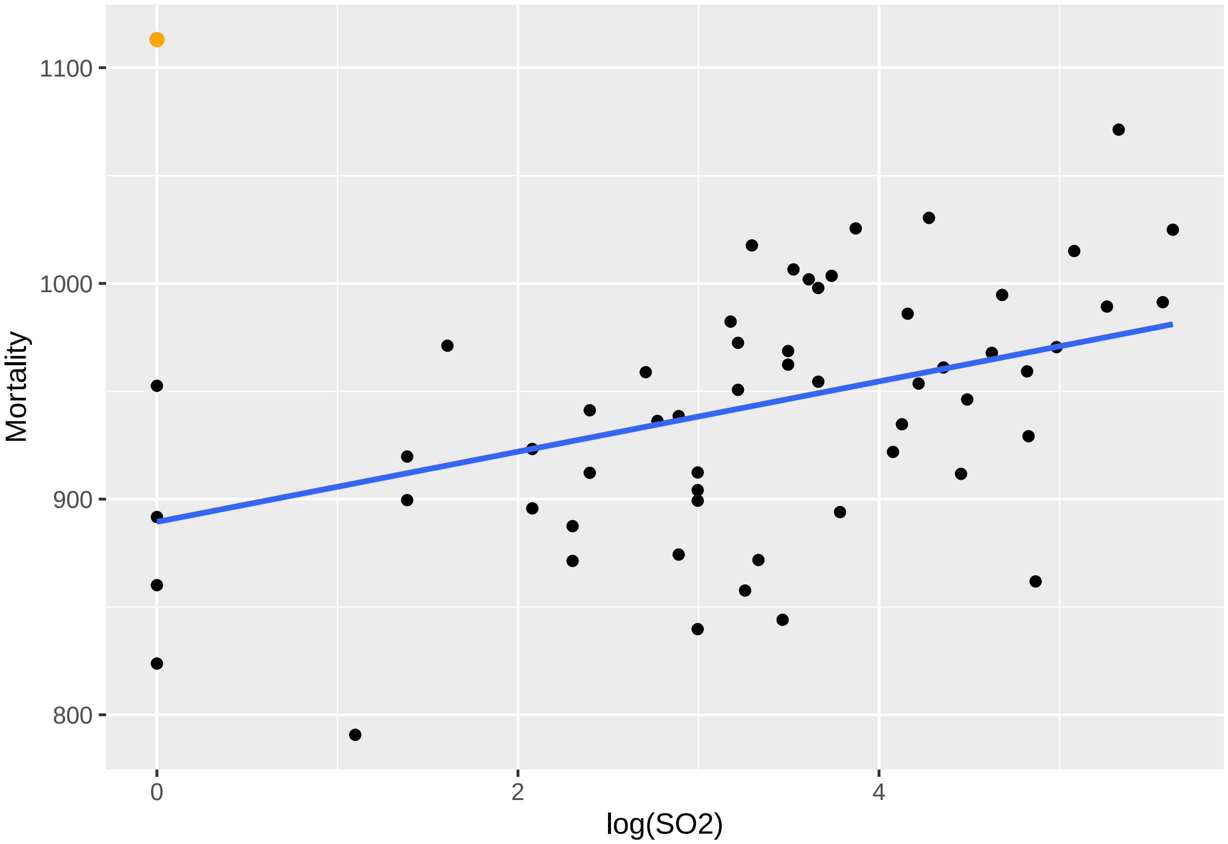
```
# case 60 eda
ggplot(without_outlier_pm, aes(log(Density), Mortality)) +
  geom_point() +
  geom_point(data=filter(without_outlier_pm, case == 60), color="orange", size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(without_outlier_pm, aes(log(SO2), Mortality)) +
  geom_point() +
  geom_point(data=filter(without_outlier_pm, case == 60), color="orange", size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

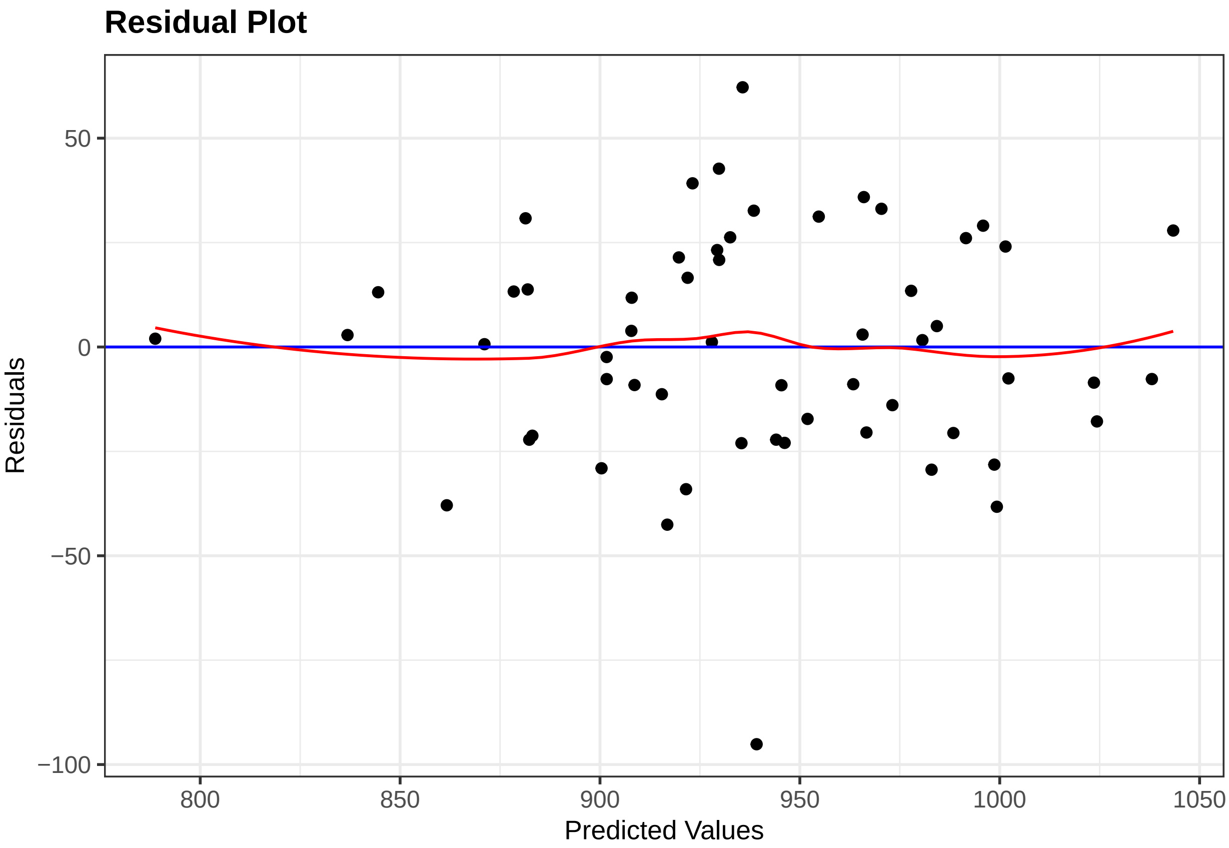


```
# slice out case 60
pm %>% slice(60)
```

```
##             CITY Mortality Precip Humidity JanTemp JulyTemp Over65 House Educ
## 1 New Orleans, LA   1113.06     54       62      54       81    7.4  3.36  9.7
##   Sound Density NonWhite WhiteCol Poor HC NOX SO2
## 1  72.8    3172     31.4     45.5 24.2 20  17   1
```

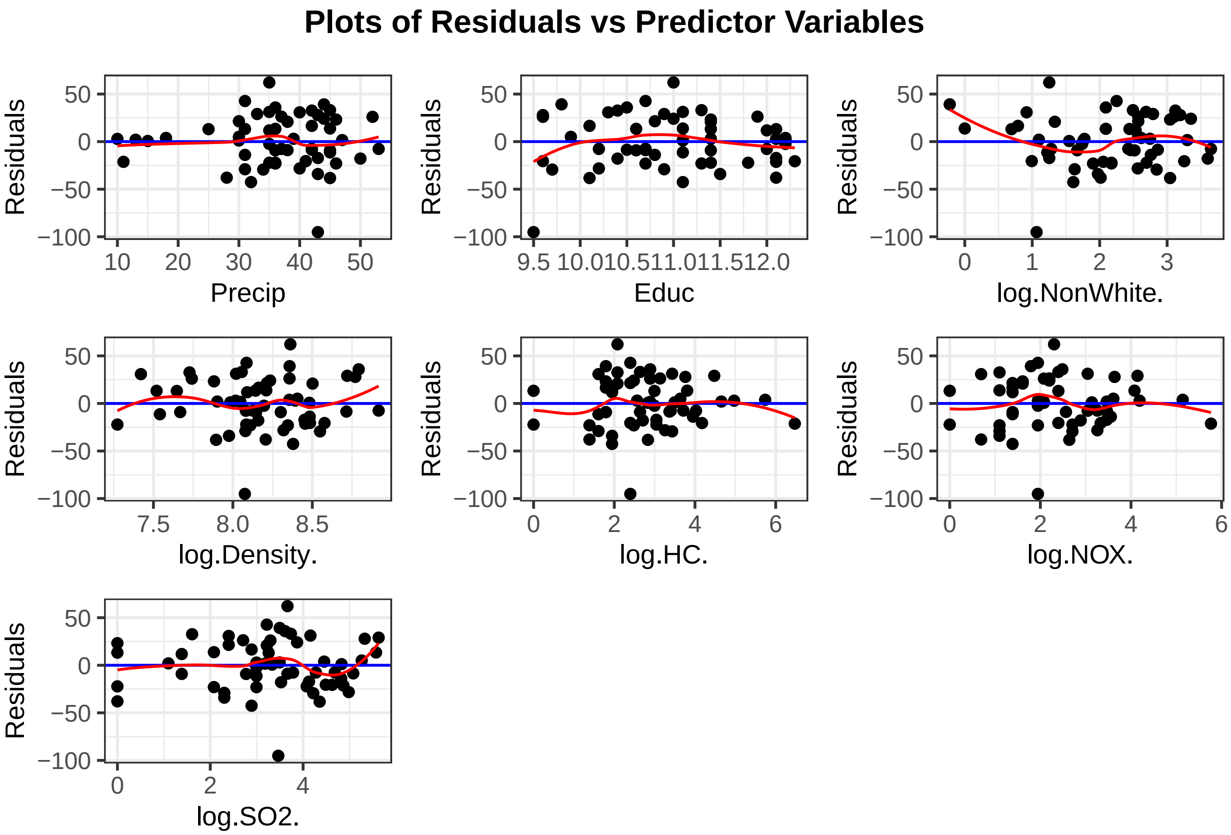
```
# check assumptions 3 - pollution
# residuals plot
resid_panel(pollution_log_lm_no_7_20_60, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

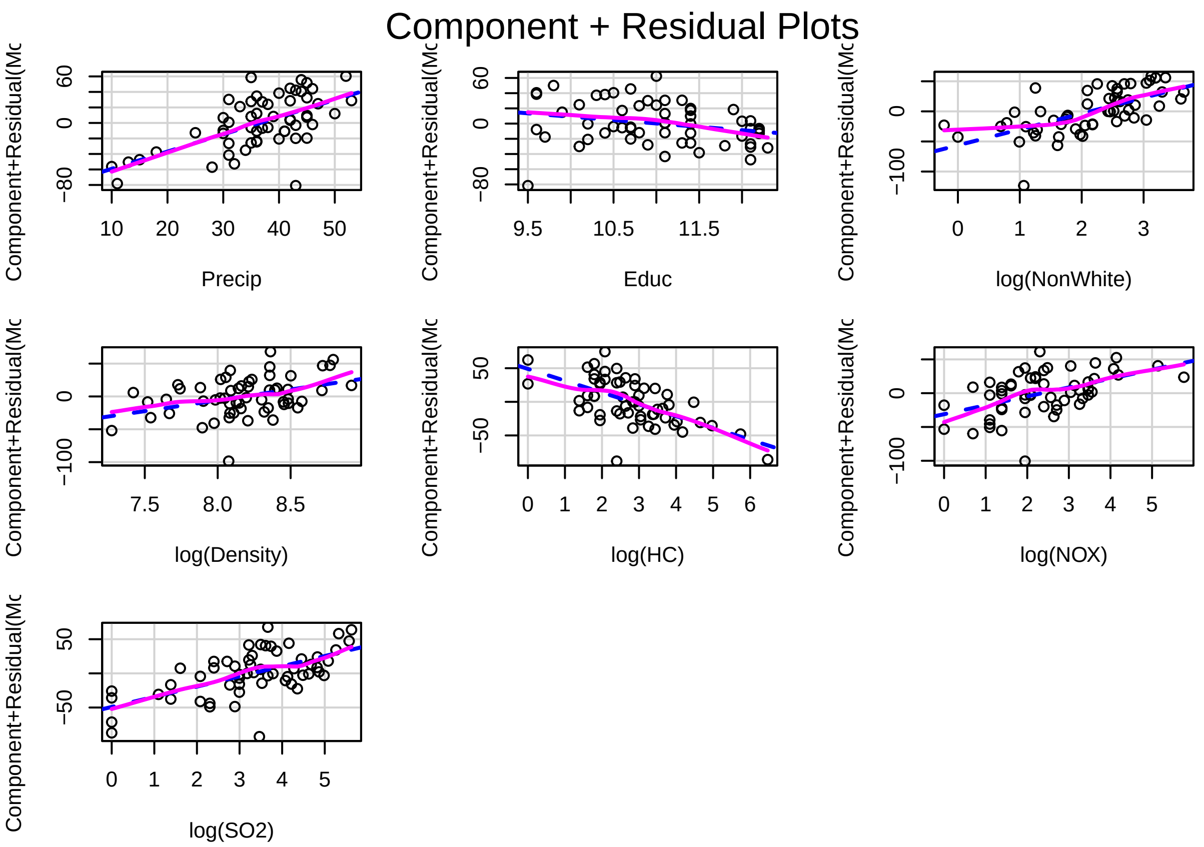


```
# residuals of each predictor
resid_xpanel(pollution_log_lm_no_7_20_60, smoother = TRUE)
```

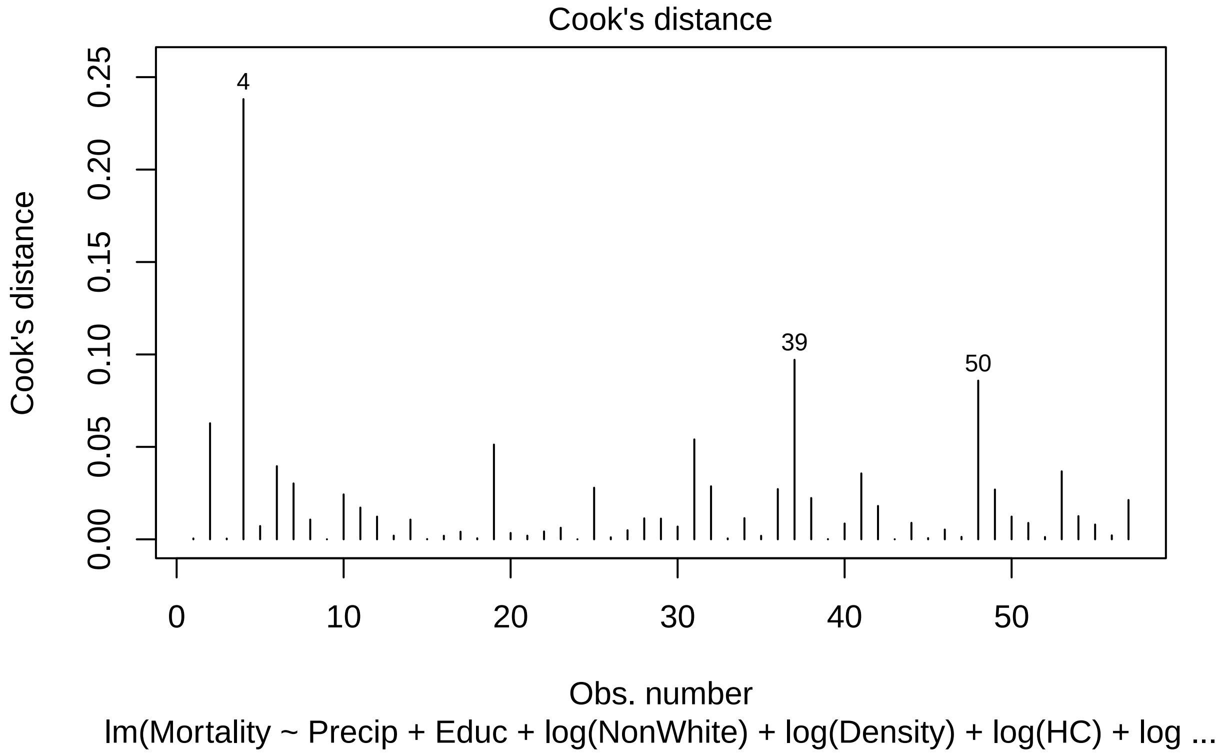
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

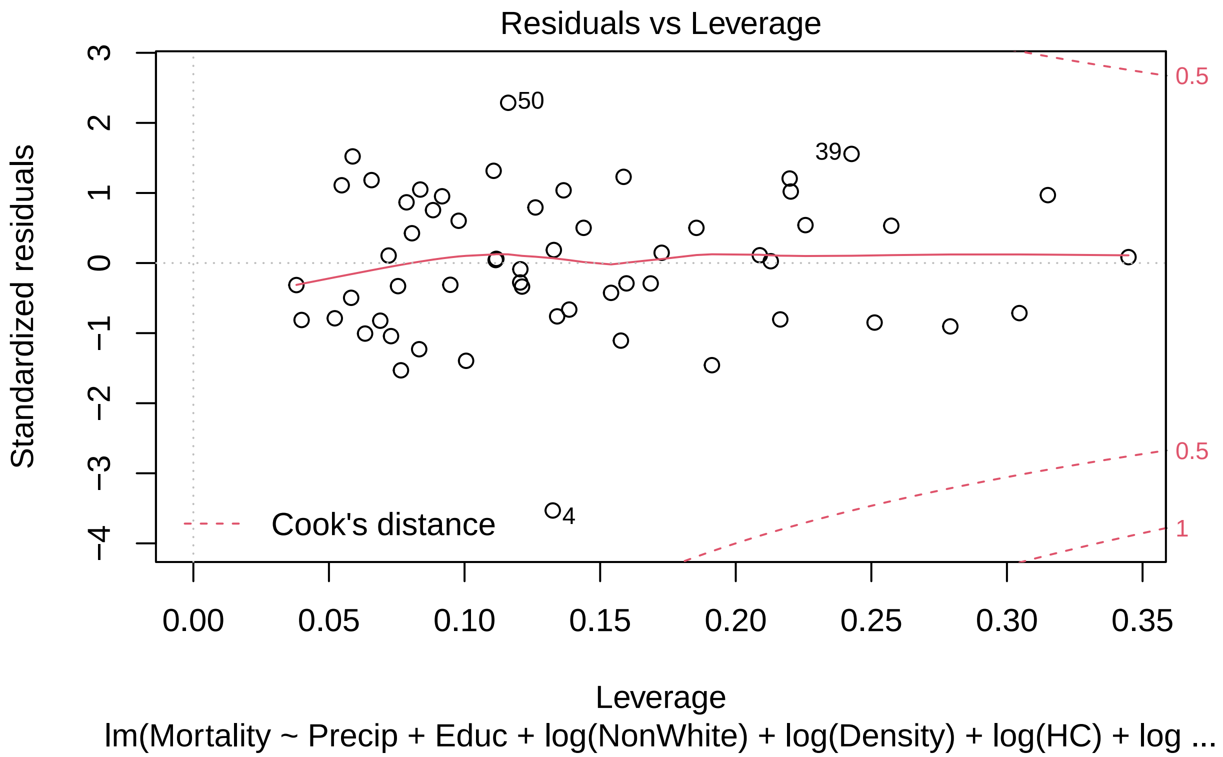


```
# partial residuals
crp(pollution_log_lm_no_7_20_60)
```



```
# check outliers 2 - pollution
plot(pollution_log_lm_no_7_20_60, which =c(4,5))
```



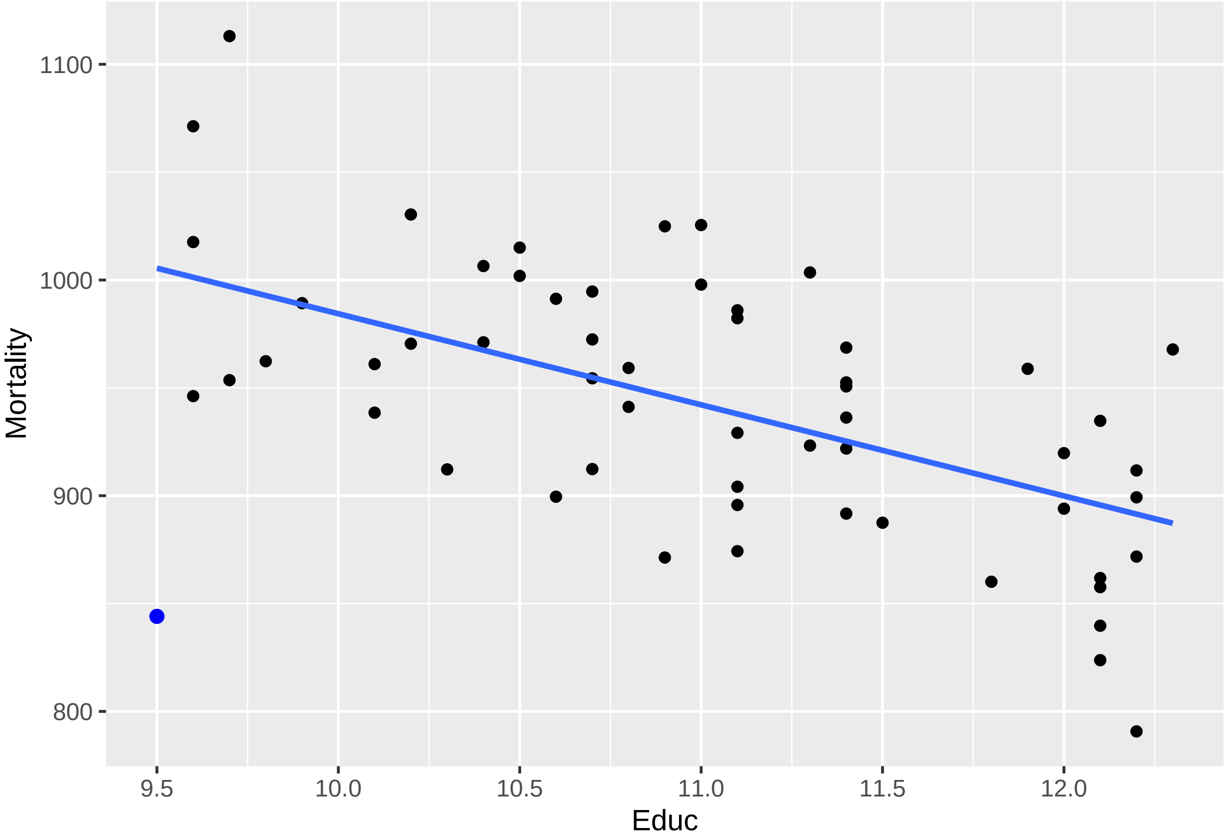


```
# refit model without case 7, case 20, case 60, and case 4
pollution_log_lm_no_7_20_60_4 <- lm(Mortality~Precip+Educ+log(NonWhite)+log(Density)+log(HC)+log(NOX)+l
# Educ is significant again, case 4 is influential
summary(pollution_log_lm_no_7_20_60_4)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(NonWhite) + log(Density) +
##     log(HC) + log(NOX) + log(SO2), data = pm, subset = -c(7,
##     20, 60, 4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.199 -17.078   0.007  18.796  60.677
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   684.6909   118.6555   5.770 5.61e-07 ***
## Precip          2.2986     0.5098   4.509 4.20e-05 ***
## Educ          -14.9890     5.7316  -2.615   0.0119 *
## log(NonWhite)  22.9126     4.3533   5.263 3.27e-06 ***
## log(Density)   31.2236    12.4547   2.507   0.0156 *
## log(HC)       -13.7032    10.1667  -1.348   0.1840
## log(NOX)       11.3911    11.1814   1.019   0.3134
## log(SO2)       13.4878     5.1265   2.631   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.24 on 48 degrees of freedom
## Multiple R-squared:  0.8321, Adjusted R-squared:  0.8077
## F-statistic: 33.99 on 7 and 48 DF,  p-value: < 2.2e-16
```

```
# case 4 eda
ggplot(without_outlier_pm, aes(Educ, Mortality)) +
  geom_point() +
  geom_point(data=filter(without_outlier_pm, case == 4), color="blue", size=2) +
  geom_smooth(method="lm", se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

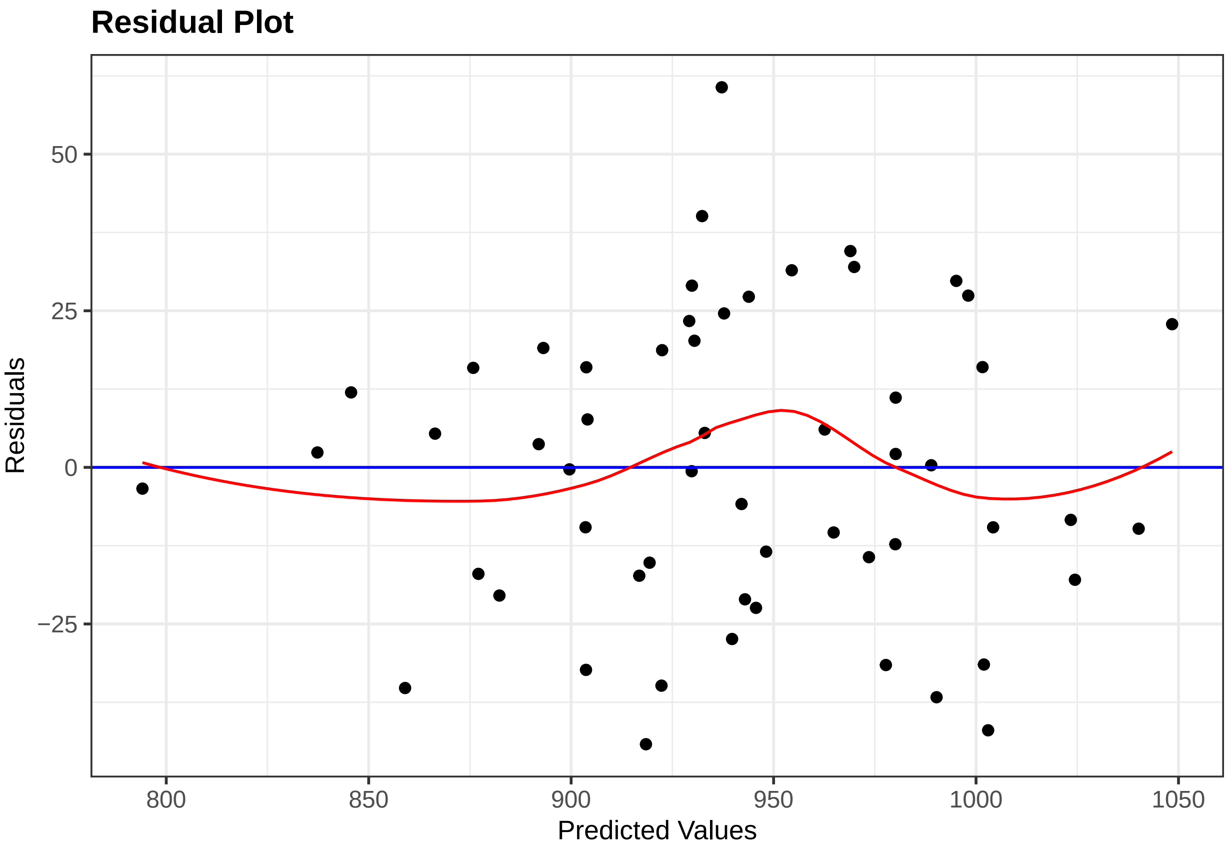


```
# slice out case 4
pm %>% slice(4)
```

```
##             CITY Mortality Precip Humidity JanTemp JulyTemp Over65 House Educ
## 1 Lancaster, PA    844.05     43       54      32       74   10.1  3.38  9.5
##   Sound Density NonWhite WhiteCol Poor HC NOX SO2
## 1  79.2    3214      2.9     43.7   12 11   7  32
```

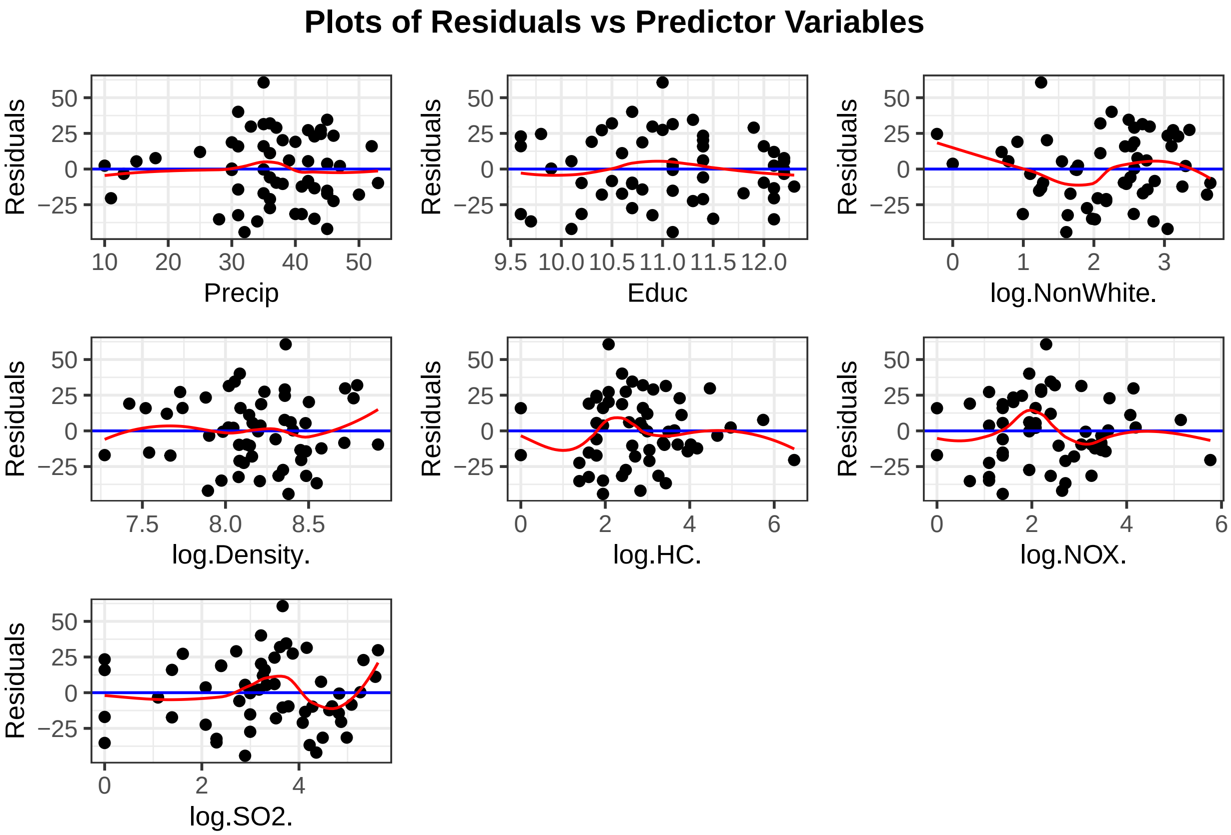
```
# check assumptions 4 - pollution
# residuals plot
resid_panel(pollution_log_lm_no_7_20_60_4, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

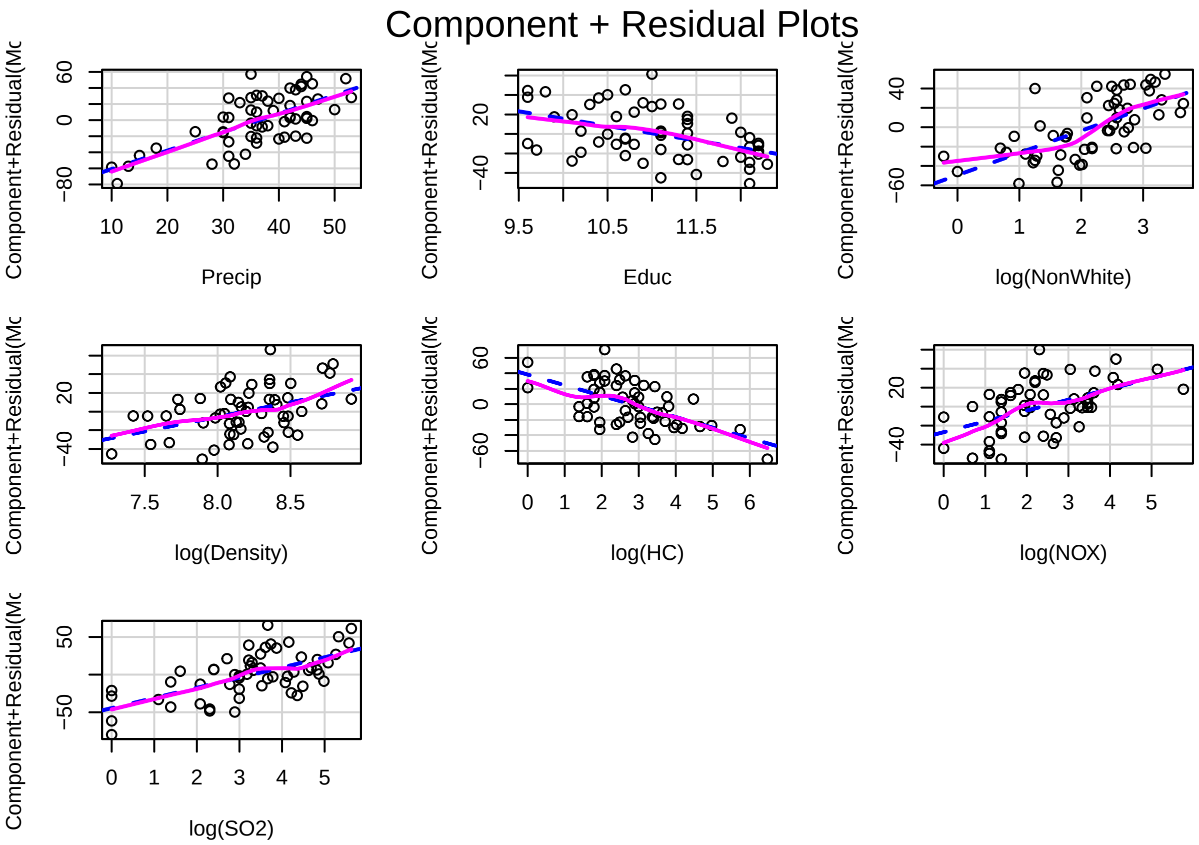


```
# residuals of each predictor
resid_xpanel(pollution_log_lm_no_7_20_60_4, smoother = TRUE)
```

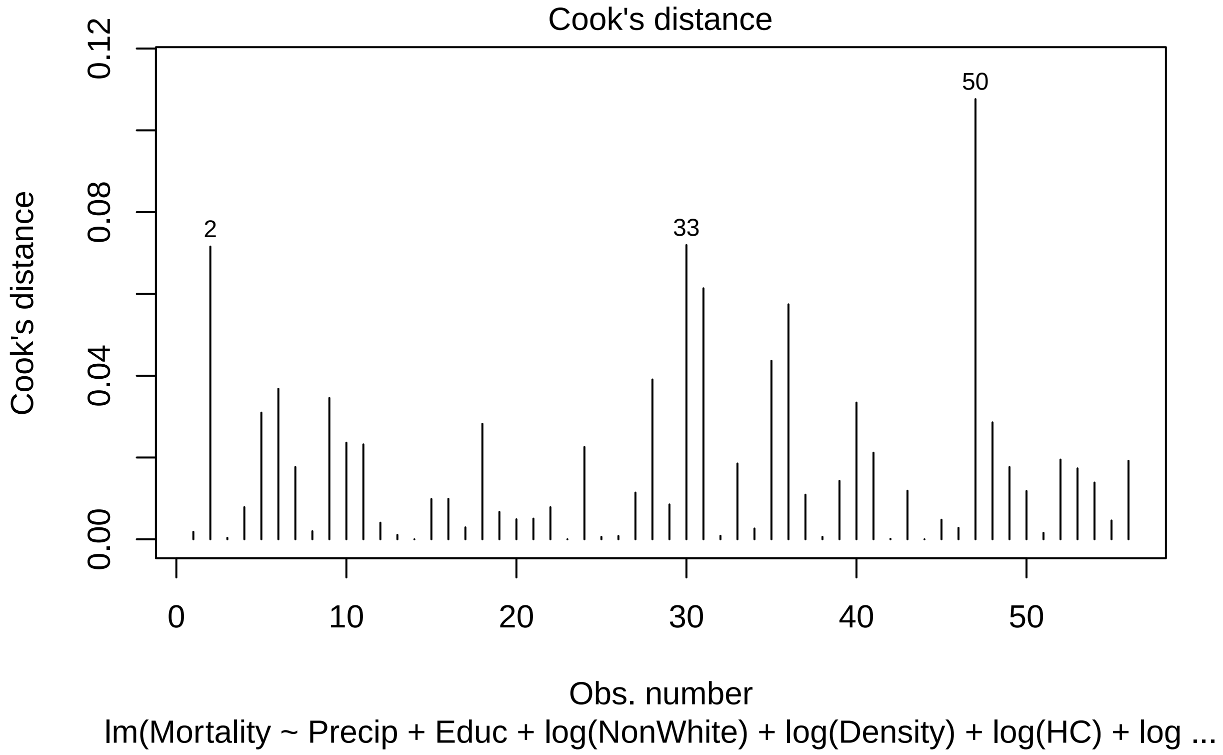
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

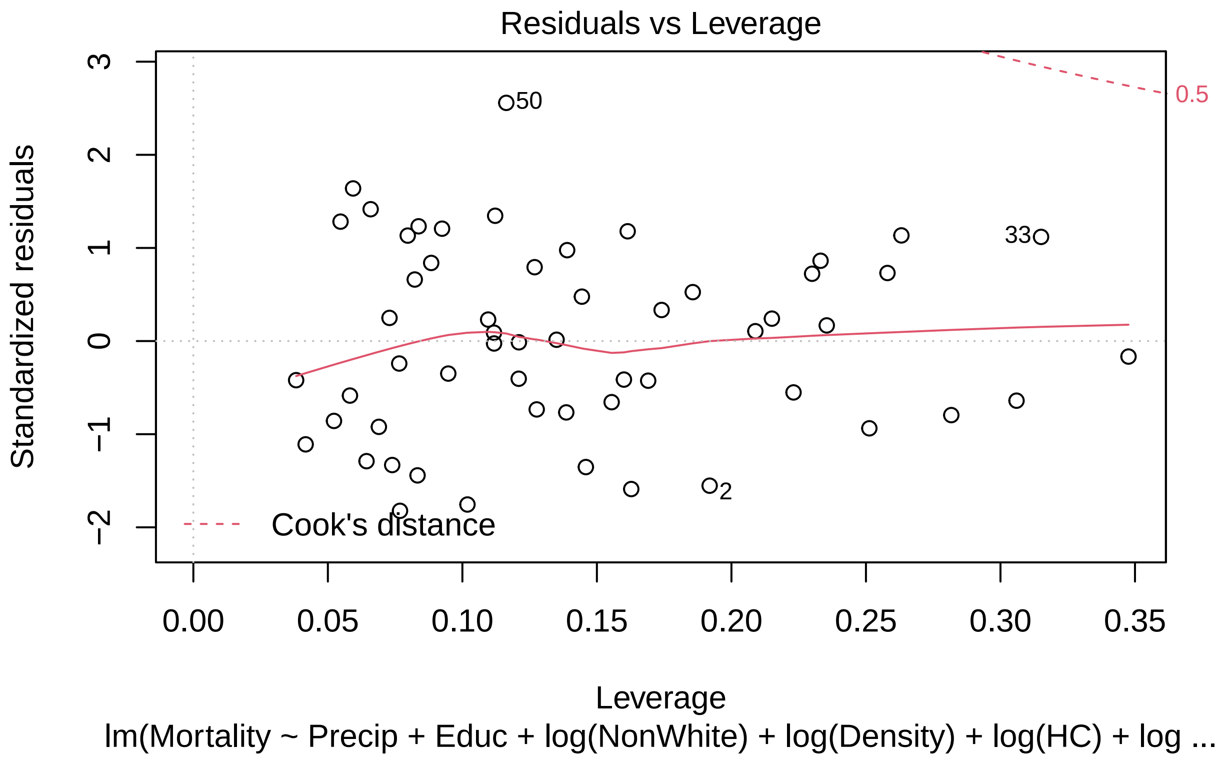


```
# partial residuals
crp(pollution_log_lm_no_7_20_60_4)
```



```r
# check outliers 3 - pollution
plot(pollution_log_lm_no_7_20_60_4, which =c(4,5))
```





```
# refit model without case 7, case 20, case 60, case 4, and case 50
pollution_log_lm_no_7_20_60_4_50 <- lm(Mortality~Precip+Educ+log(NonWhite)+log(Density)+log(HC)+log(NOX)
# case 50 is not influential
summary(pollution_log_lm_no_7_20_60_4_50)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(NonWhite) + log(Density) +
##     log(HC) + log(NOX) + log(SO2), data = pm, subset = -c(7,
##     20, 60, 4, 50))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.549 -15.690  -1.701  17.016  41.489
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   701.0211   111.6018   6.281 1.01e-07 ***
## Precip          2.3941     0.4801   4.987 8.79e-06 ***
## Educ          -15.6755     5.3890  -2.909  0.00553 **
## log(NonWhite)  24.3183     4.1211   5.901 3.80e-07 ***
## log(Density)   28.6140    11.7366   2.438  0.01860 *
## log(HC)        -6.3983     9.9182  -0.645  0.52199
## log(NOX)        5.0810    10.7541   0.472  0.63878
## log(SO2)       13.2941     4.8153   2.761  0.00820 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.7 on 47 degrees of freedom
## Multiple R-squared:  0.8523, Adjusted R-squared:  0.8303
## F-statistic: 38.76 on 7 and 47 DF,  p-value: < 2.2e-16
```

```
# anova 1 - pollution
smaller_lm1 <- lm(Mortality~Precip+Educ+log(NonWhite)+log(Density)+log(SO2), data=pm, subset=-c(7, 20, 
bigger_lm1 <- lm(Mortality~Precip+Educ+log(NonWhite)+log(Density)+log(HC)+log(NOX)+log(SO2), data=pm, su
anova(smaller_lm1, bigger_lm1)
```

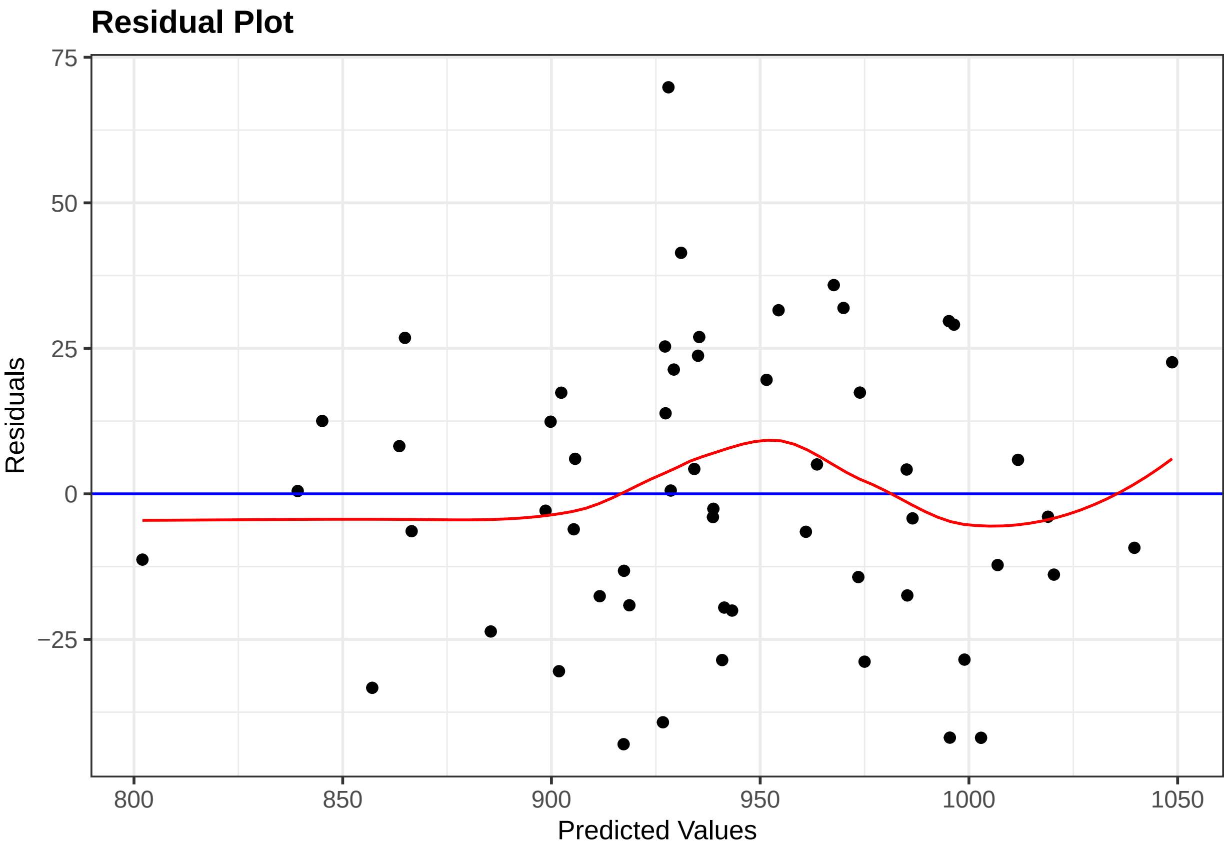
```
## Analysis of Variance Table
##
## Model 1: Mortality ~ Precip + Educ + log(NonWhite) + log(Density) + log(SO2)
## Model 2: Mortality ~ Precip + Educ + log(NonWhite) + log(Density) + log(HC) +
##     log(NOX) + log(SO2)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     50 31762
## 2     48 30570  2    1192.5 0.9362 0.3991
```

```
# every term is significant
summary(smaller_lm1)
```

```
##
## Call:
## lm(formula = Mortality ~ Precip + Educ + log(NonWhite) + log(Density) +
##     log(SO2), data = pm, subset = -c(7, 20, 60, 4))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43.01 -17.48  -2.73  17.95  69.85
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   692.4902   118.1558   5.861 3.60e-07 ***
## Precip          2.5249     0.4358   5.794 4.57e-07 ***
## Educ          -16.5763     5.4713  -3.030 0.003869 **
## log(NonWhite)  22.3121     4.1014   5.440 1.60e-06 ***
## log(Density)   30.4036    12.4216   2.448 0.017934 *
## log(SO2)       12.9177     3.2353   3.993 0.000214 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.2 on 50 degrees of freedom
## Multiple R-squared:  0.8256, Adjusted R-squared:  0.8082
## F-statistic: 47.34 on 5 and 50 DF,  p-value: < 2.2e-16
```

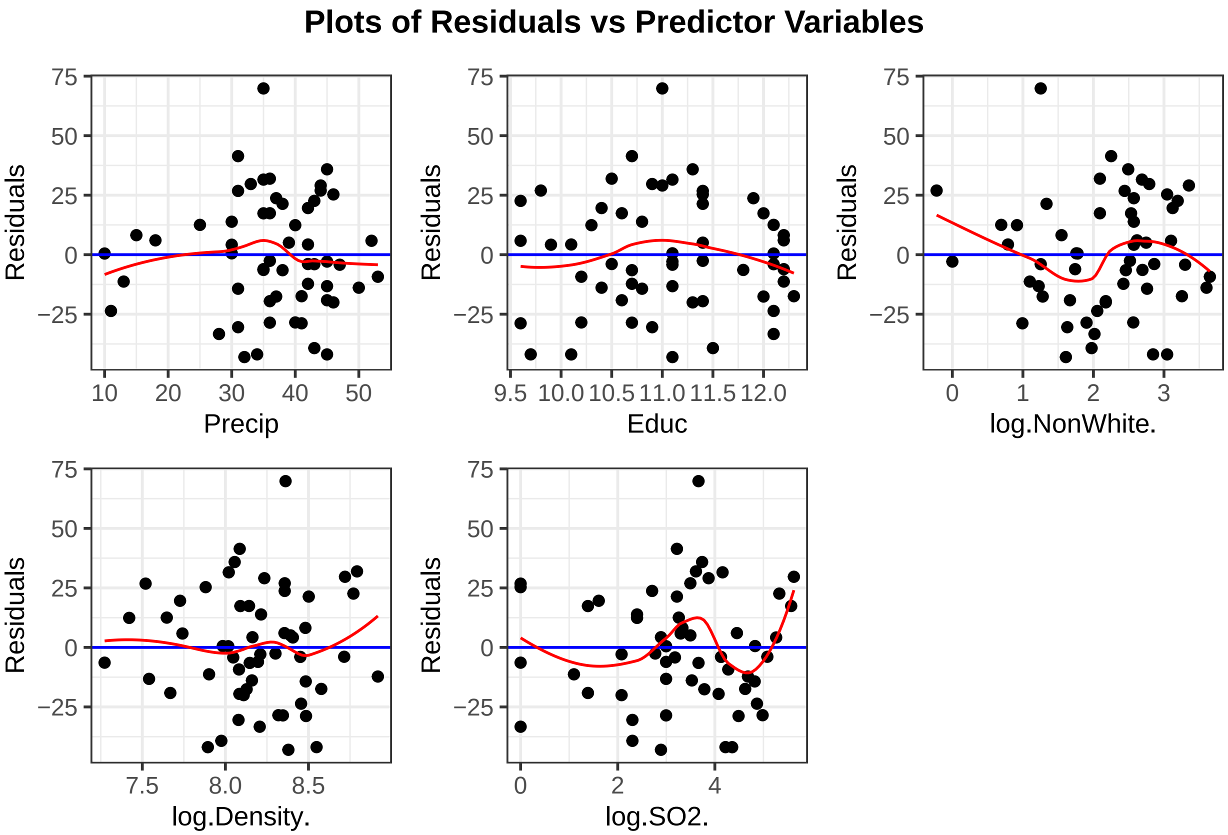
```
# check assumptions 5 - pollution
# residuals plot
resid_panel(smaller_lm1, plots = "resid", smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

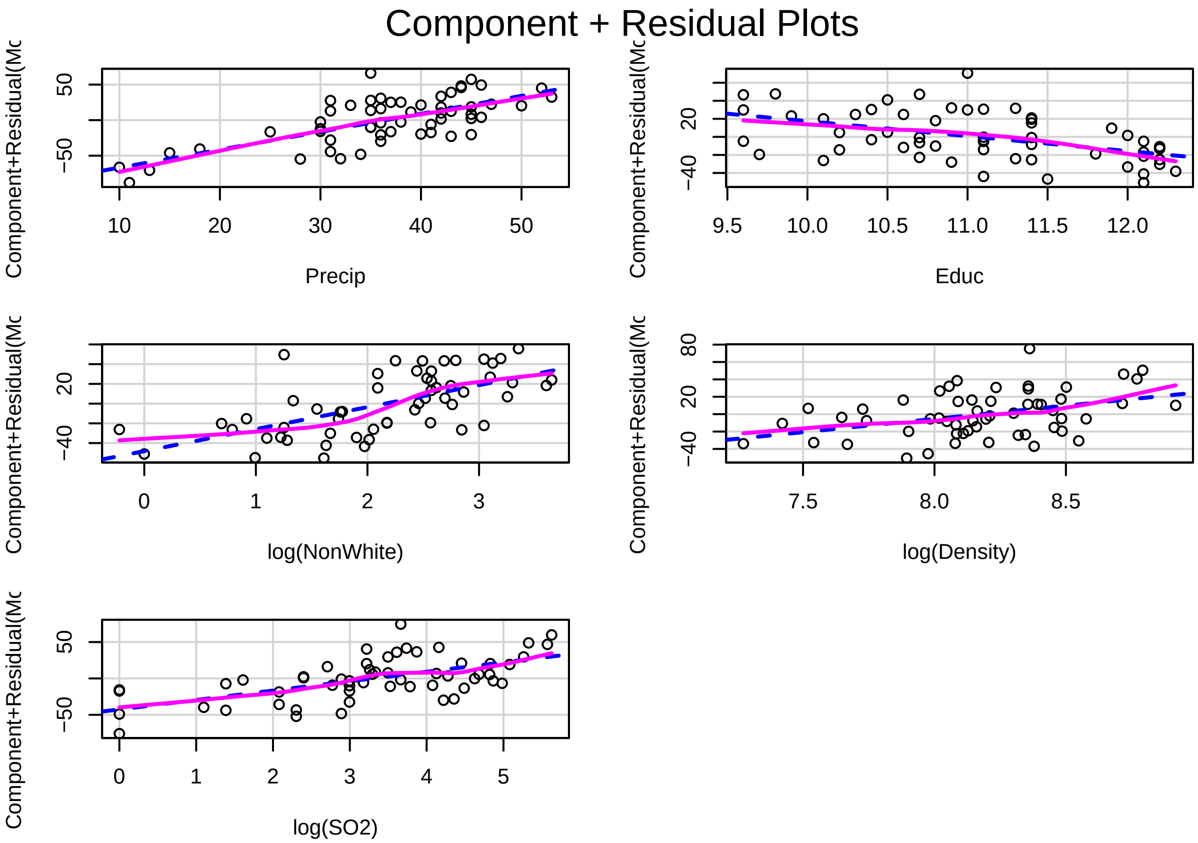


```
# residuals of each predictor
resid_xpanel(smaller_lm1, smoother = TRUE)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



```
# partial residuals
crp(smaller_lm1)
```



```r
# check collinearity 2 - pollution
vif(smaller_lm1)
```

```
##        Precip          Educ log(NonWhite)  log(Density)       log(SO2)
##      1.493798      1.596570      1.086489      1.615862      1.825687
```