STAT 331 Final Project

Maxine, Estella, Judy, Weiwei

04/12/2021

Requirement of the project

Your 7–10 page report must contain the following components:

- 1. Summary: A maximum of 200 words describing the objective of the report, an overview of the statistical analysis, and summary of the main results.
- 2. Objective: Describe your goals for the analysis.
- 3. Exploratory Data Analysis: Conduct exploratory data analyses: report summary statistics, visualize data (histograms, scatter plots, etc.). Report on any interesting findings and comment on how these inform the rest of your analysis.
- 4. Methods: Describe your statistical analysis: What is your model? Did you use any transformations or extensions of the basic multiple linear regression model? How did you select a model? Does the model fit the data well? Are the necessary assumptions met? Be sure to explain and justify your decisions.
- 5. Results: Report on the findings of your analysis
- 6. Discussion: Comment on your findings/conclusions; describe any limitations of your analysis.

1. Summary

A maximum of 200 words describing the objective of the report, an overview of the statistical analysis, and summary of the main results.

2. Objective

The goal of this project is to analyze the pollutants.csv data and write a report on your analysis. The specific goals of your analysis are up to you to decide.

3. Exploratory Data Analysis

Conduct exploratory data analyses: report summary statistics, visualize data (histograms, scatter plots, etc.). Report on any interesting findings and comment on how these inform the rest of your analysis.

can use this as a tutorial https://r4ds.had.co.nz/exploratory-data-analysis.html

Take a peak at the first 5 entries

```
# CHANGE ABSOLUTE PATH
# pollutants <- read.csv("~/R331project/data/pollutants.csv")
pollutants <- read.csv("~/Desktop/R331project/pollutants.csv")
head(pollutants)
```

```
##   X    length POP_PCB1 POP_PCB2 POP_PCB3 POP_PCB4 POP_PCB5 POP_PCB6 POP_PCB7
## 1 1 1.1587651    20000     7600     3700    14700    18900     5300     5500
## 2 2 0.9011283    43900    14900     9700    32300    55500    13400    18700
## 3 3 1.2753948     3300     3300     3300     3300     3300     3300     3300
## 4 4 0.9369063     8500     4100     6000    11500    13500     6900    13500
## 5 5 0.7027998   159000    60200    29800   170000   215000    79200    47400
## 6 6 1.1516147    14400     7100    16900    28200    37200    22000    10200
##   POP_PCB8 POP_PCB9 POP_PCB10 POP_PCB11 POP_dioxin1 POP_dioxin2 POP_dioxin3
## 1     5700     2000      15.6      23.1        70.9        50.0         173
## 2    12000    16200      35.4      31.1       116.0       129.0         709
## 3     3300     3300       1.8       9.3        29.9         5.4         148
## 4     4100     4100       4.5      21.1        50.4        29.4         668
## 5    41400    53900      59.2      80.3        98.1        80.1         875
## 6     3800     6400      19.2      70.0       106.0        47.4         533
##   POP_furan1 POP_furan2 POP_furan3 POP_furan4 whitecell_count lymphocyte_pct
## 1        6.9        5.6        0.8       15.6             5.4           33.8
## 2       18.5       15.4       20.3        2.3             5.6           16.8
## 3        1.3        1.4        1.2        2.9             6.3           35.3
## 4        2.2        2.4        2.3       43.2             8.4           23.0
## 5       13.7        1.2        0.8       11.0             6.7           24.5
## 6        8.3        7.0        3.4       19.4             4.7           39.5
##   monocyte_pct eosinophils_pct basophils_pct neutrophils_pct   BMI edu_cat
## 1          8.1            51.2           6.2             0.6 27.50       2
## 2         10.2            69.4           3.2             0.5 27.46       3
## 3          7.3            54.9           1.6             0.9 36.13       1
## 4          6.4            68.8           1.7             0.2 21.79       4
## 5          7.5            64.3           3.0             0.8 31.46       2
## 6          4.4            54.2           1.3             0.8 40.68       1
##   race_cat male ageyrs yrssmoke smokenow ln_lbxcot
## 1        4    1     41        0        0 -2.312635
## 2        4    0     77        0        0 -4.509860
## 3        2    0     22        0        0 -4.017384
## 4        4    0     27        0        0 -3.863233
## 5        4    1     78        0        0 -1.826351
## 6        3    0     35        0        0 -2.207275
```

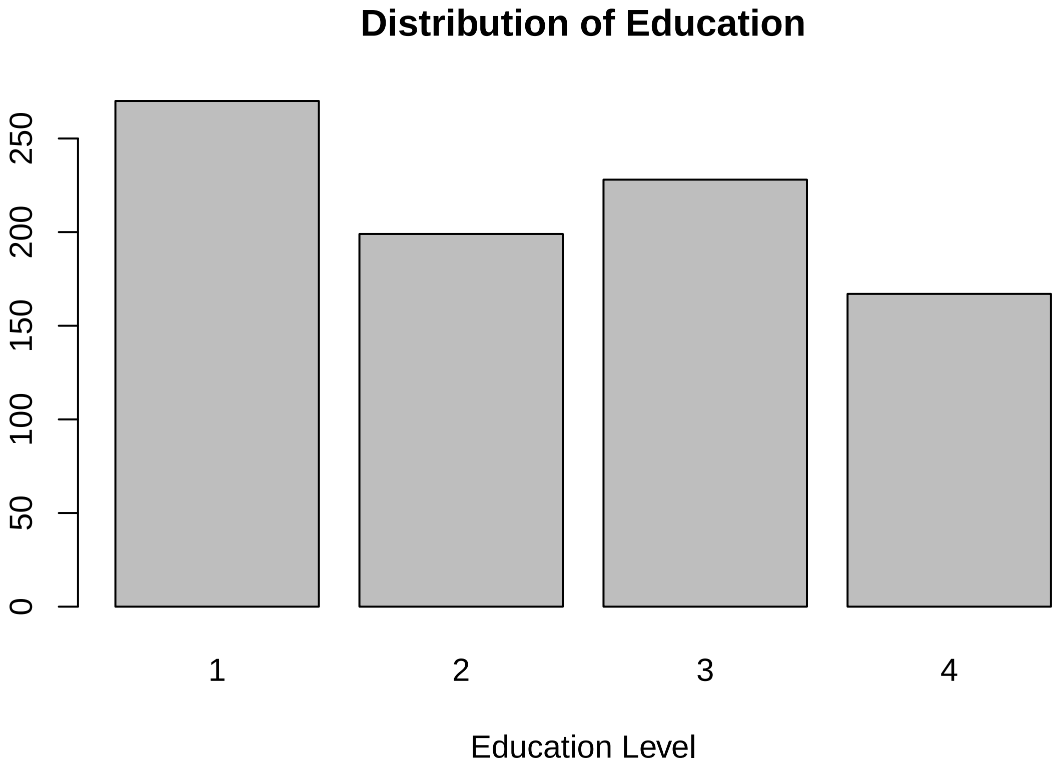
Covariates

```
names(pollutants)
```

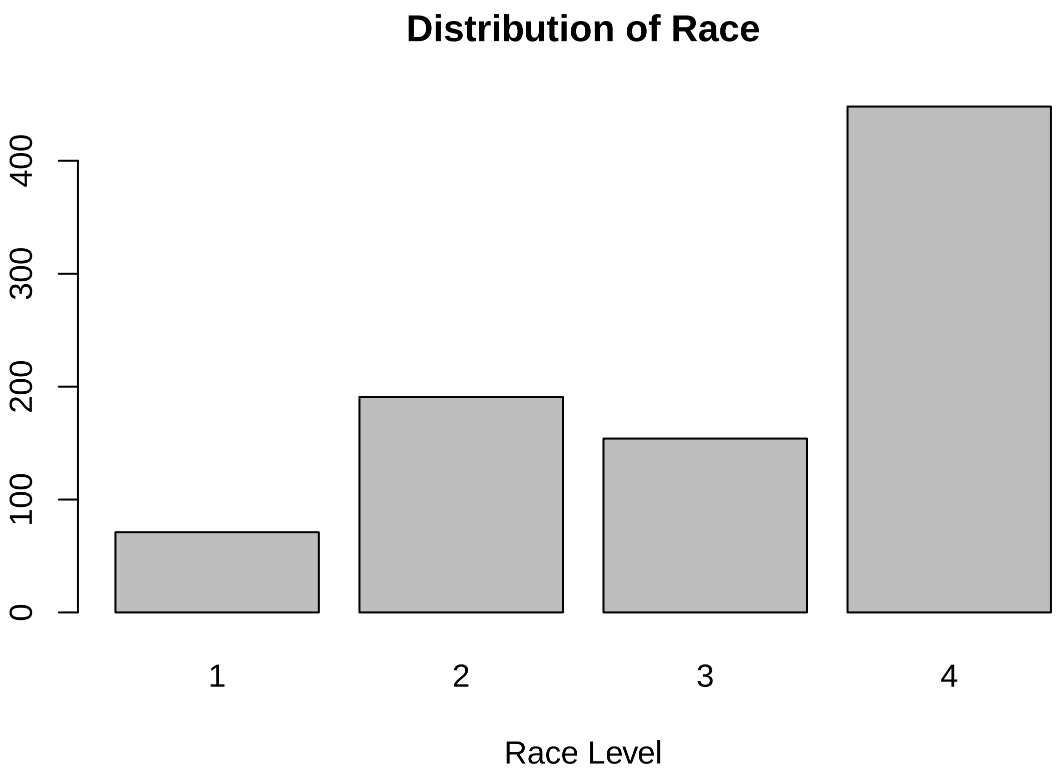
```
##  [1] "X"               "length"          "POP_PCB1"        "POP_PCB2"
##  [5] "POP_PCB3"        "POP_PCB4"        "POP_PCB5"        "POP_PCB6"
##  [9] "POP_PCB7"        "POP_PCB8"        "POP_PCB9"        "POP_PCB10"
## [13] "POP_PCB11"       "POP_dioxin1"     "POP_dioxin2"     "POP_dioxin3"
## [17] "POP_furan1"      "POP_furan2"      "POP_furan3"      "POP_furan4"
## [21] "whitecell_count" "lymphocyte_pct"  "monocyte_pct"    "eosinophils_pct"
## [25] "basophils_pct"   "neutrophils_pct" "BMI"             "edu_cat"
## [29] "race_cat"        "male"            "ageyrs"          "yrssmoke"
## [33] "smokenow"        "ln_lbxcot"
```

Note that “edu_cat”, “race_cat”, “male”, “smokenow” are categorical data.

```
# 1 = Less Than 9th Grade or 9-11th Grade (Includes 12th grade with no diploma)
# 2 = High School Grad/GED or Equivalent
# 3 = Some College or AA degree
# 4 = College Graduate

edu_factor=as.factor(pollutants$edu_cat)
plot(edu_factor,
     main="Distribution of Education",
     xlab="Education Level")
```

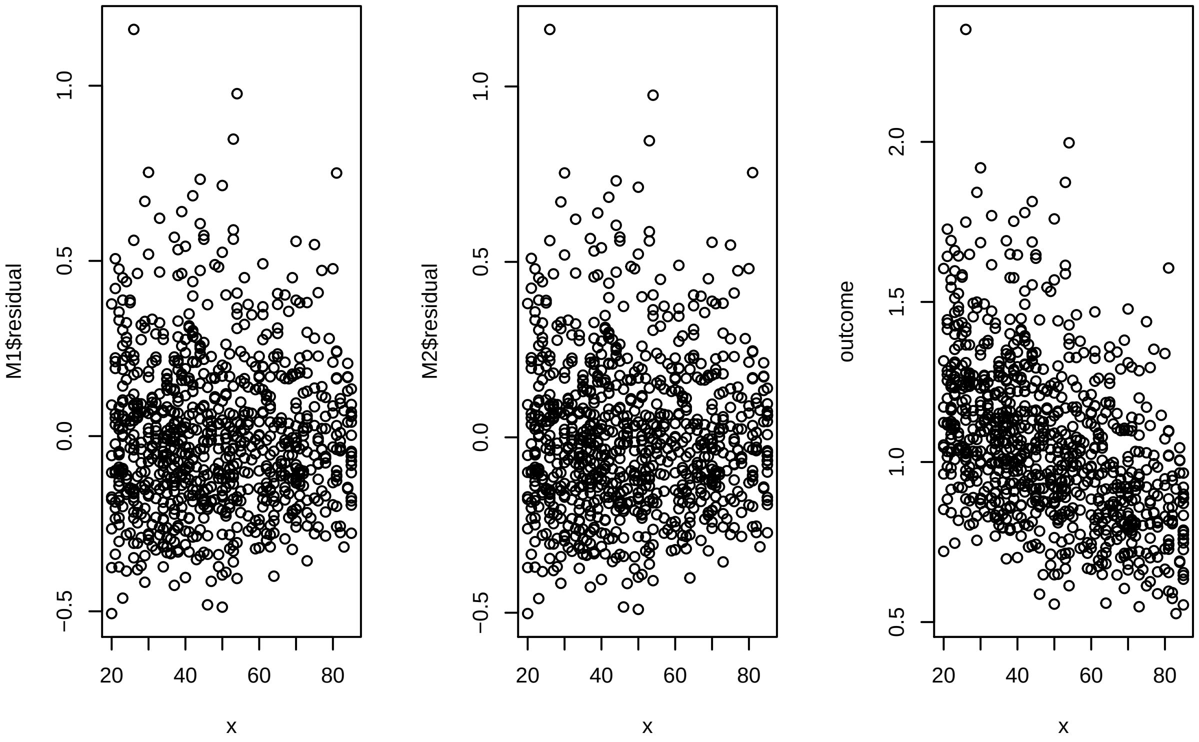


```
# 1 = Other Race (Including Multi-Racial);
# 2 = Mexican American;
# 3 = Non-Hispanic Black;
# 4 = Non-Hispanic White

race_factor=as.factor(pollutants$race_cat)
plot(race_factor,
     main="Distribution of Race",
     xlab="Race Level")
```

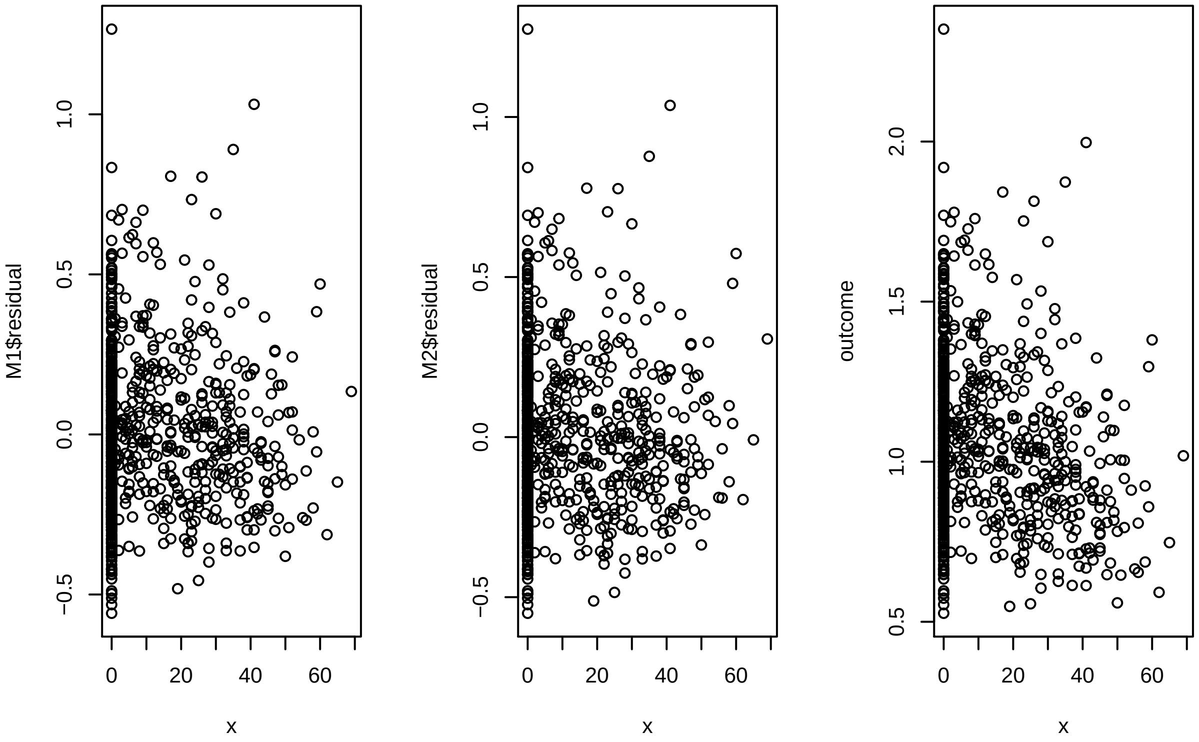


```r
# Judy's work Part 1
# testing non-linearity in SLR
# if for any covariate, residual vs x for M1 has a pattern and
# residual vs x for M2 seems random, then y has a nonlinear
# relationship with with x.
# M1: fitting y to x
# M2: fitting y to x^2

par(mfrow=c(1, 3))
outcome <- pollutants$length
check <- function(x) {
  M1 <- lm(outcome ~ x)
  print(paste("residual for M1: ", sigma(M1)))
  M2 <- lm(outcome ~ x + I(x^2))
  print(paste("residual for M2: ", sigma(M2)))
  plot(x, M1$residual)
  plot(x, M2$residual)
  plot(x, outcome)
}
list <- list(pollutants$ageyrs, pollutants$yrssmoke,
             pollutants$BMI, pollutants$ln_lbxcot,
             pollutants$whitecell_count, pollutants$lymphocyte_pct,
             pollutants$monocyte_pct, pollutants$eosinophils_pct,
             pollutants$basophils_pct, pollutants$neutrophils_pct)
for (column in list) {
  check(column)
}
```

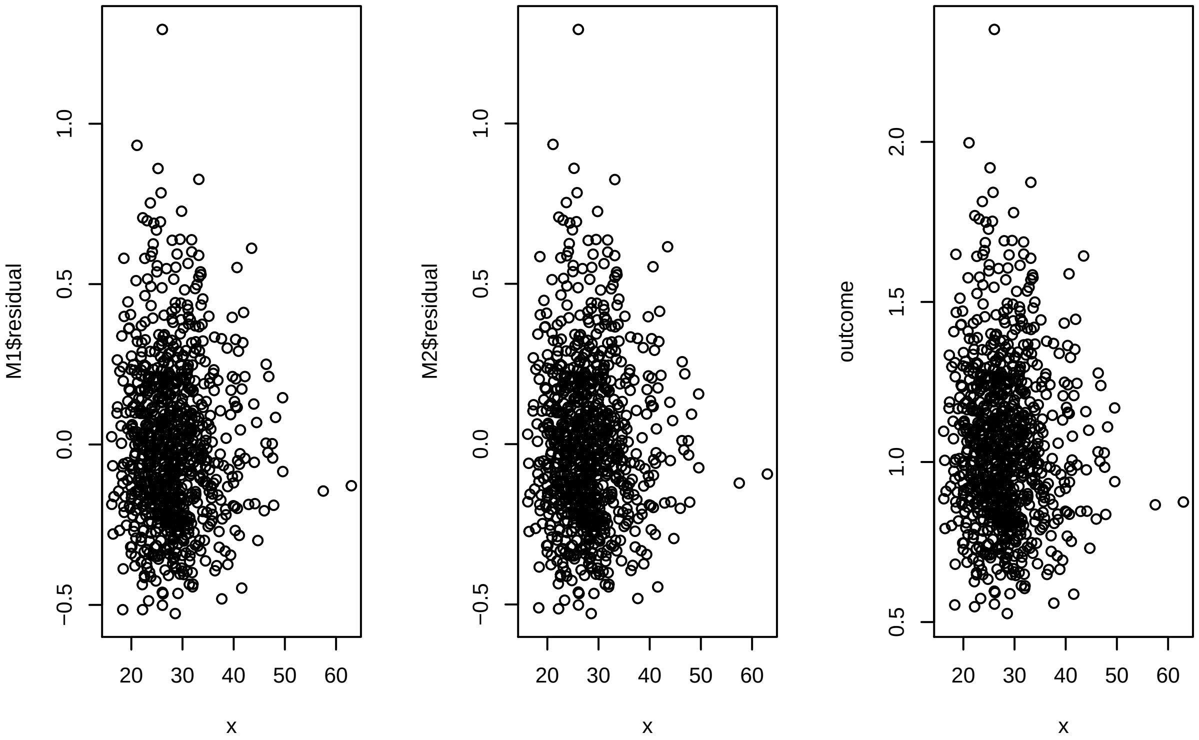
```
## [1] "residual for M1:  0.224172364185412"
## [1] "residual for M2:  0.22429269961392"
```



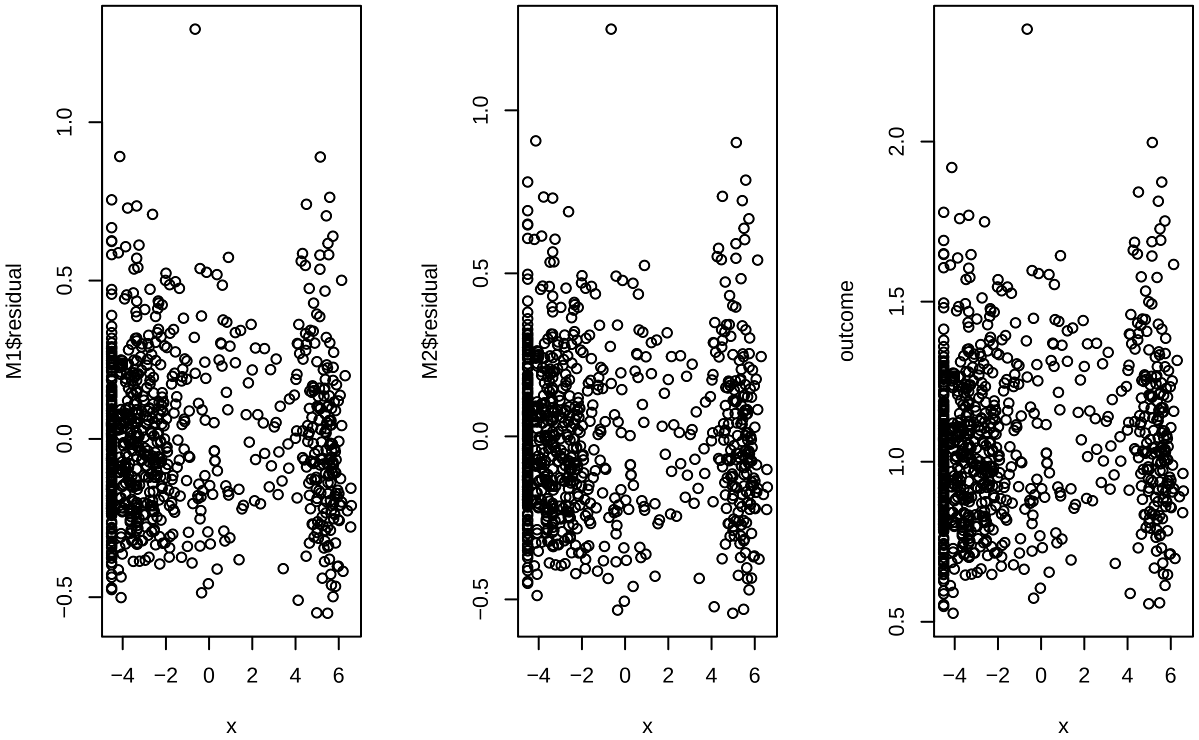
```
## [1] "residual for M1:  0.246320733146214"
## [1] "residual for M2:  0.245622720856213"
```



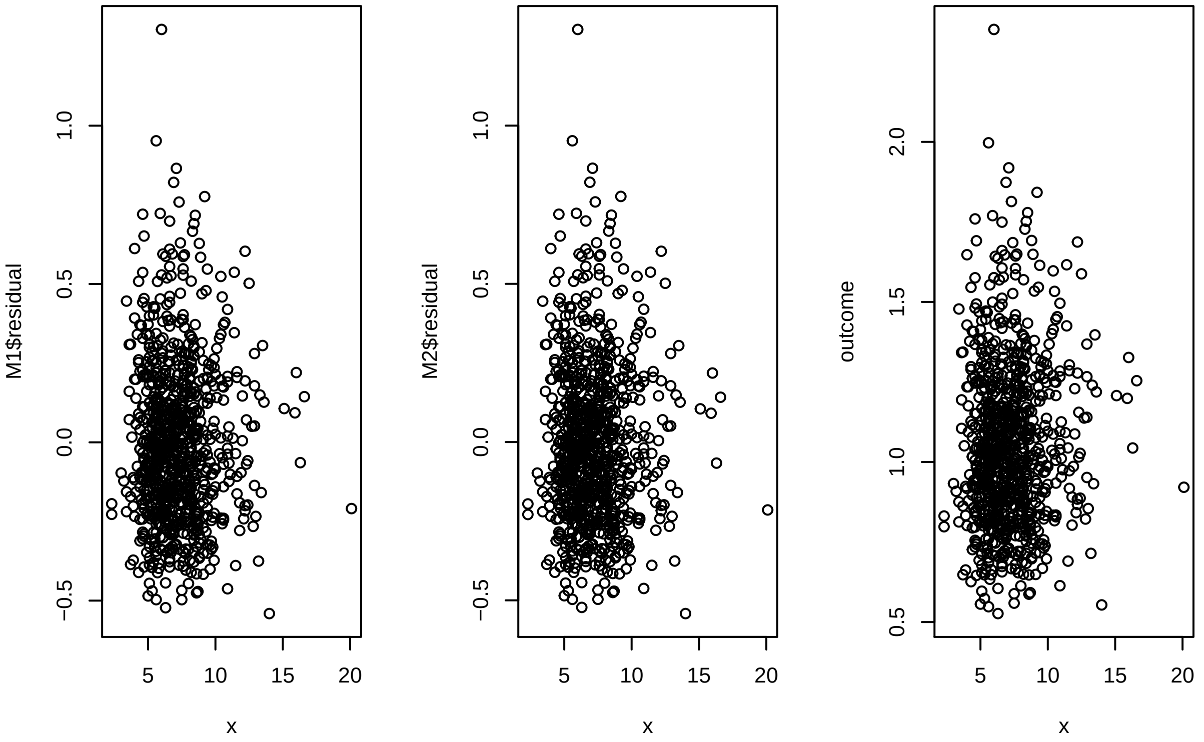
```
## [1] "residual for M1:  0.250228706427173"
## [1] "residual for M2:  0.25036248052387"
```



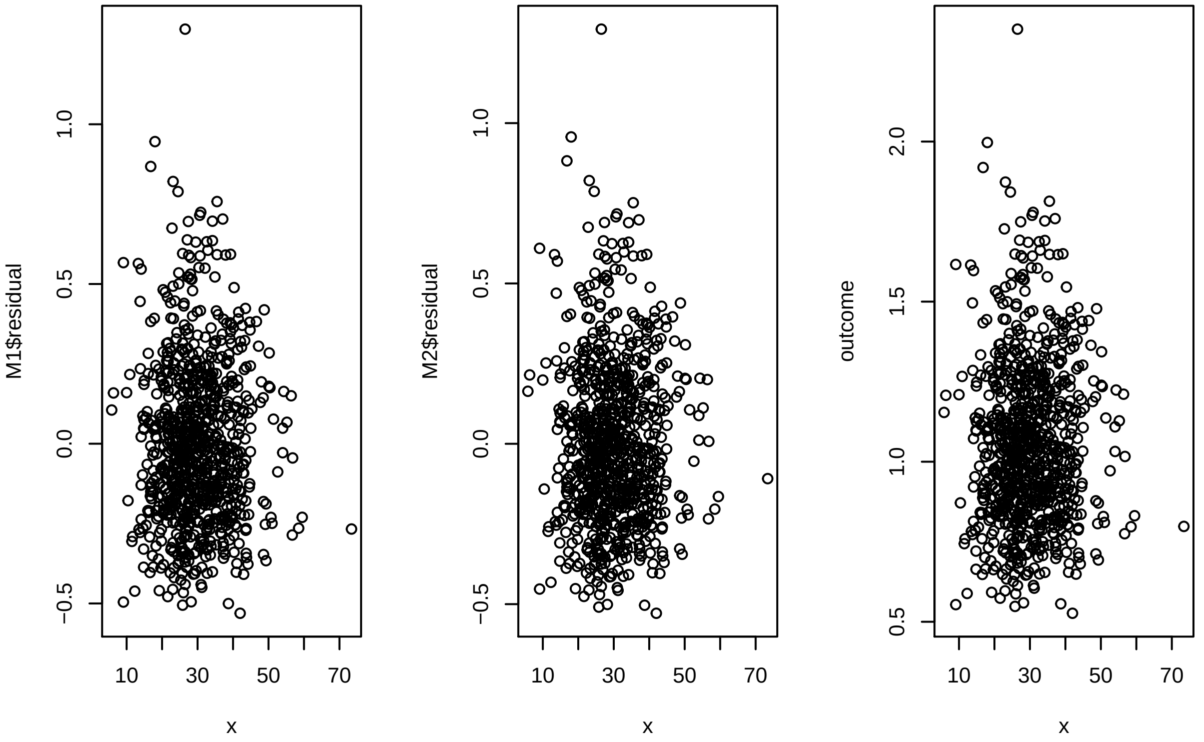
```
## [1] "residual for M1:  0.248212063673837"
## [1] "residual for M2:  0.24710732733351"
```



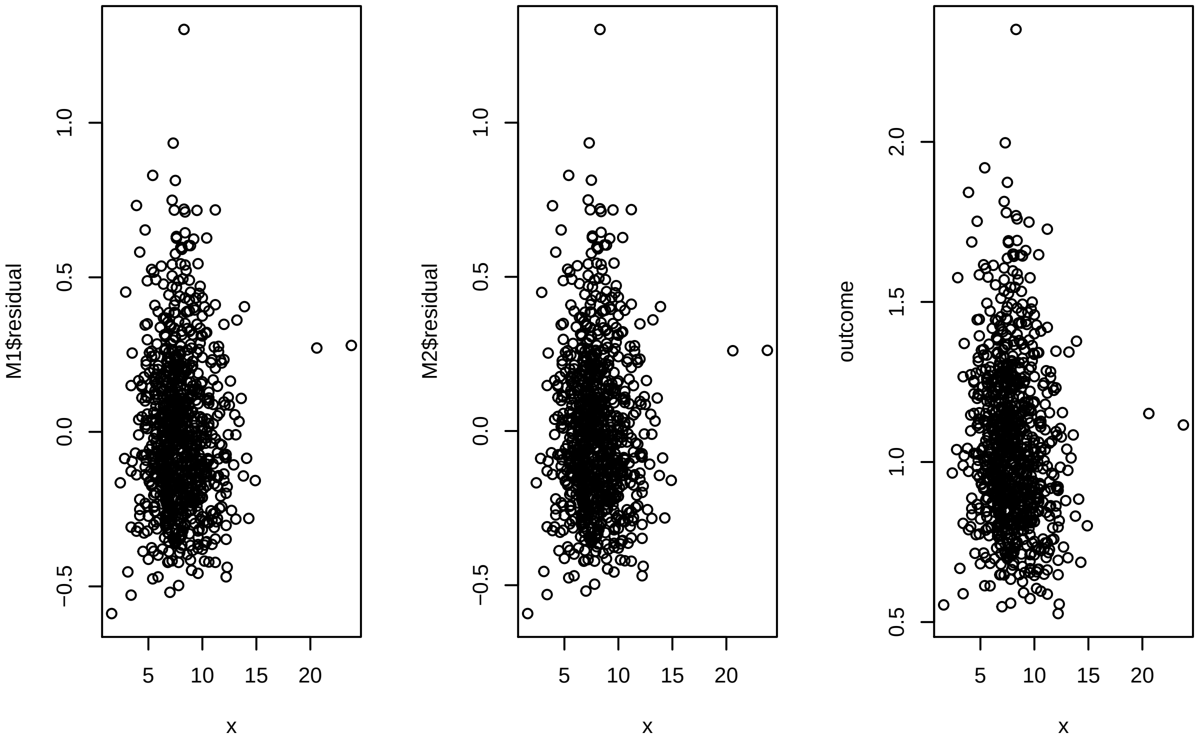
```
## [1] "residual for M1:  0.250065445847753"
## [1] "residual for M2:  0.250210403543218"
```



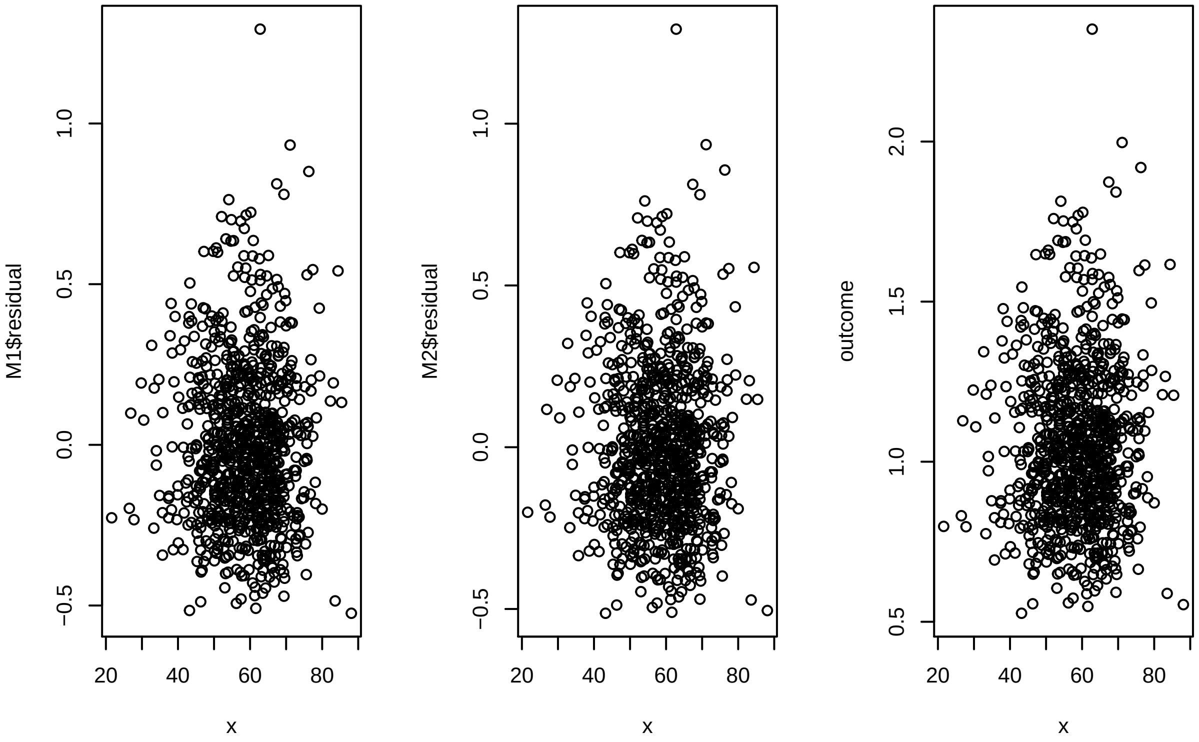
```
## [1] "residual for M1:  0.250373616826691"
## [1] "residual for M2:  0.250255208638358"
```



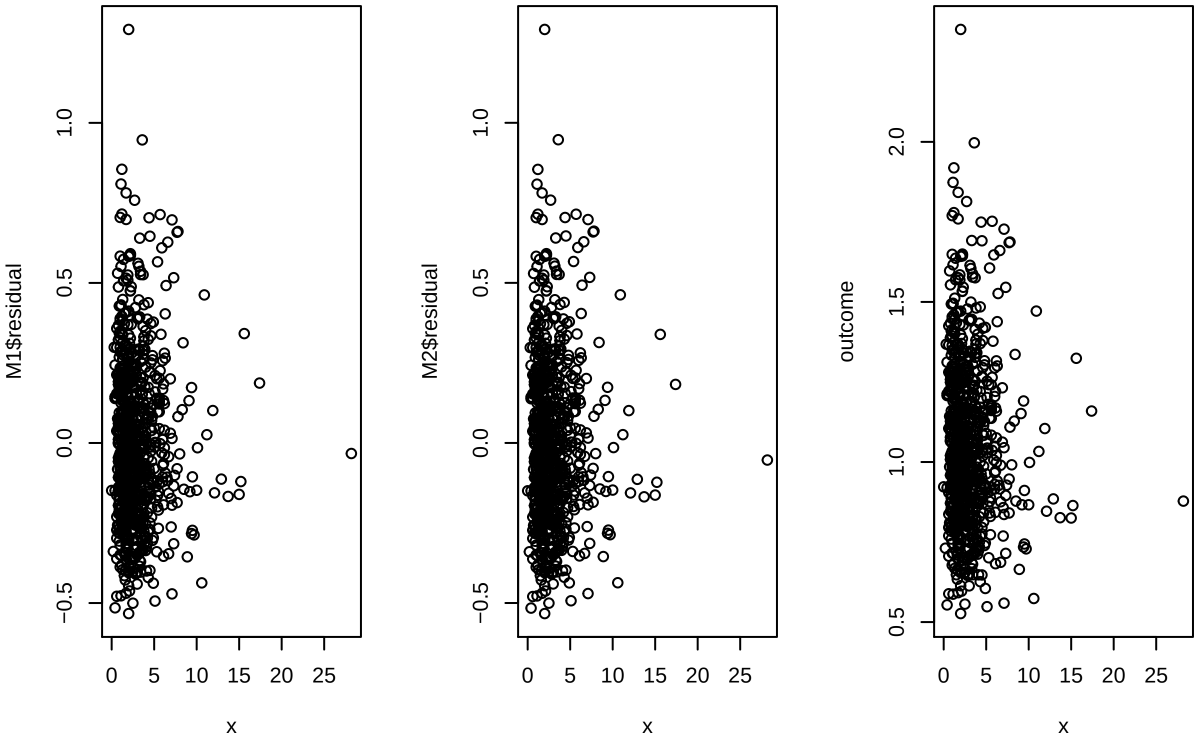
```
## [1] "residual for M1:  0.248704466454944"
## [1] "residual for M2:  0.248847192837983"
```



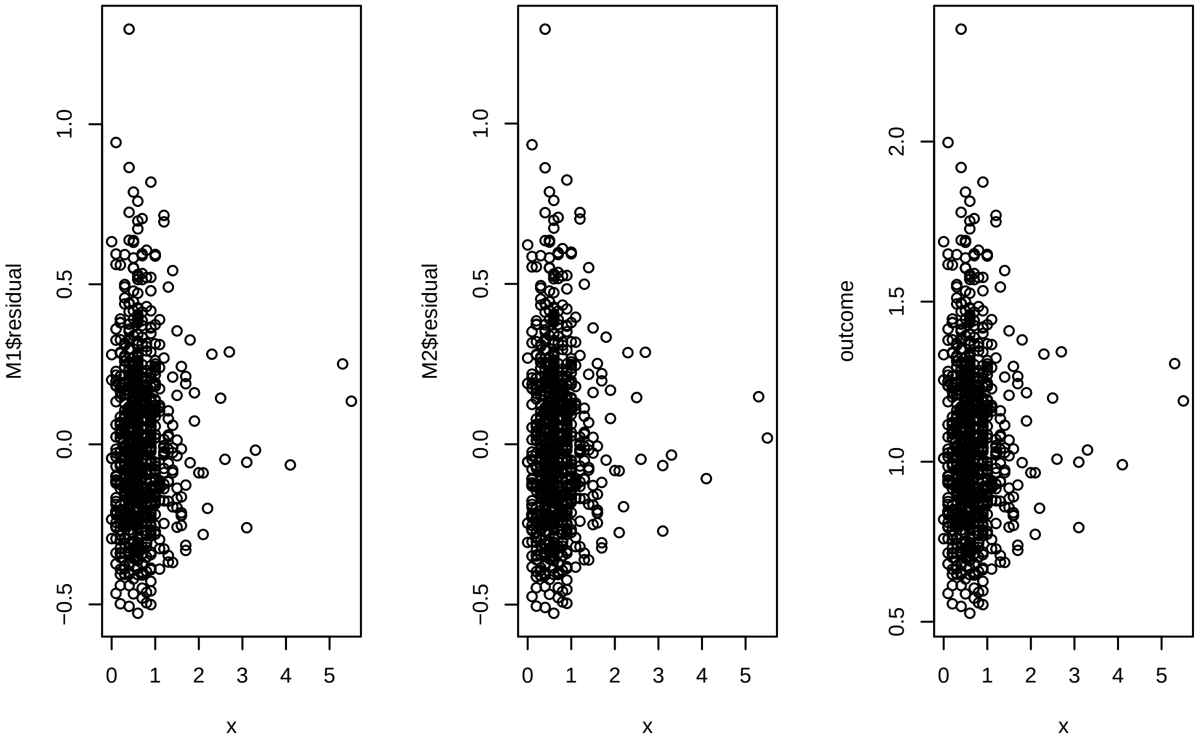
```
## [1] "residual for M1:  0.25026710930793"
## [1] "residual for M2:  0.250393729526099"
```



```
## [1] "residual for M1:  0.250043388210667"
## [1] "residual for M2:  0.25018695270193"
```



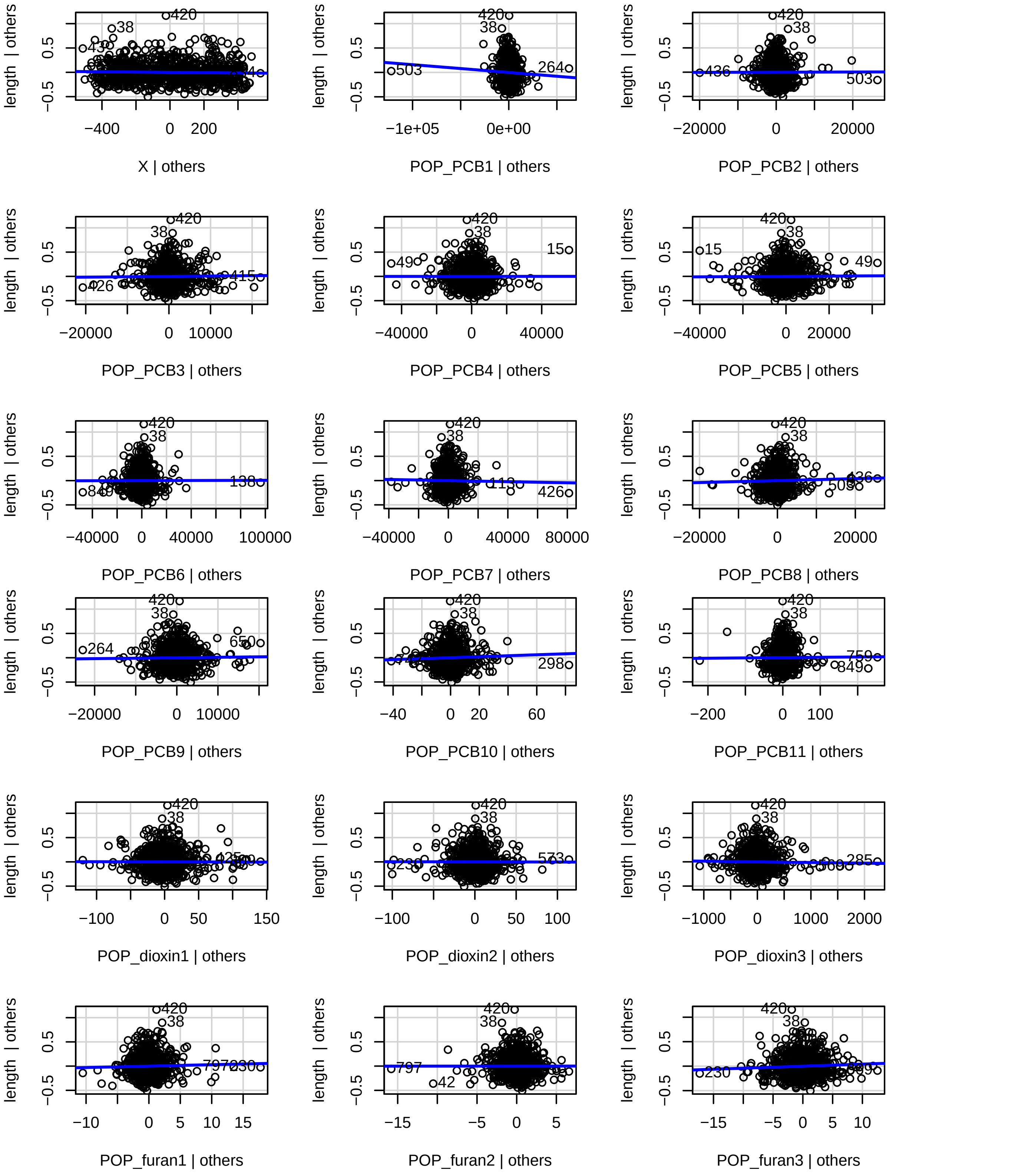
```
## [1] "residual for M1:  0.250382476371691"
## [1] "residual for M2:  0.25042580861039"
```

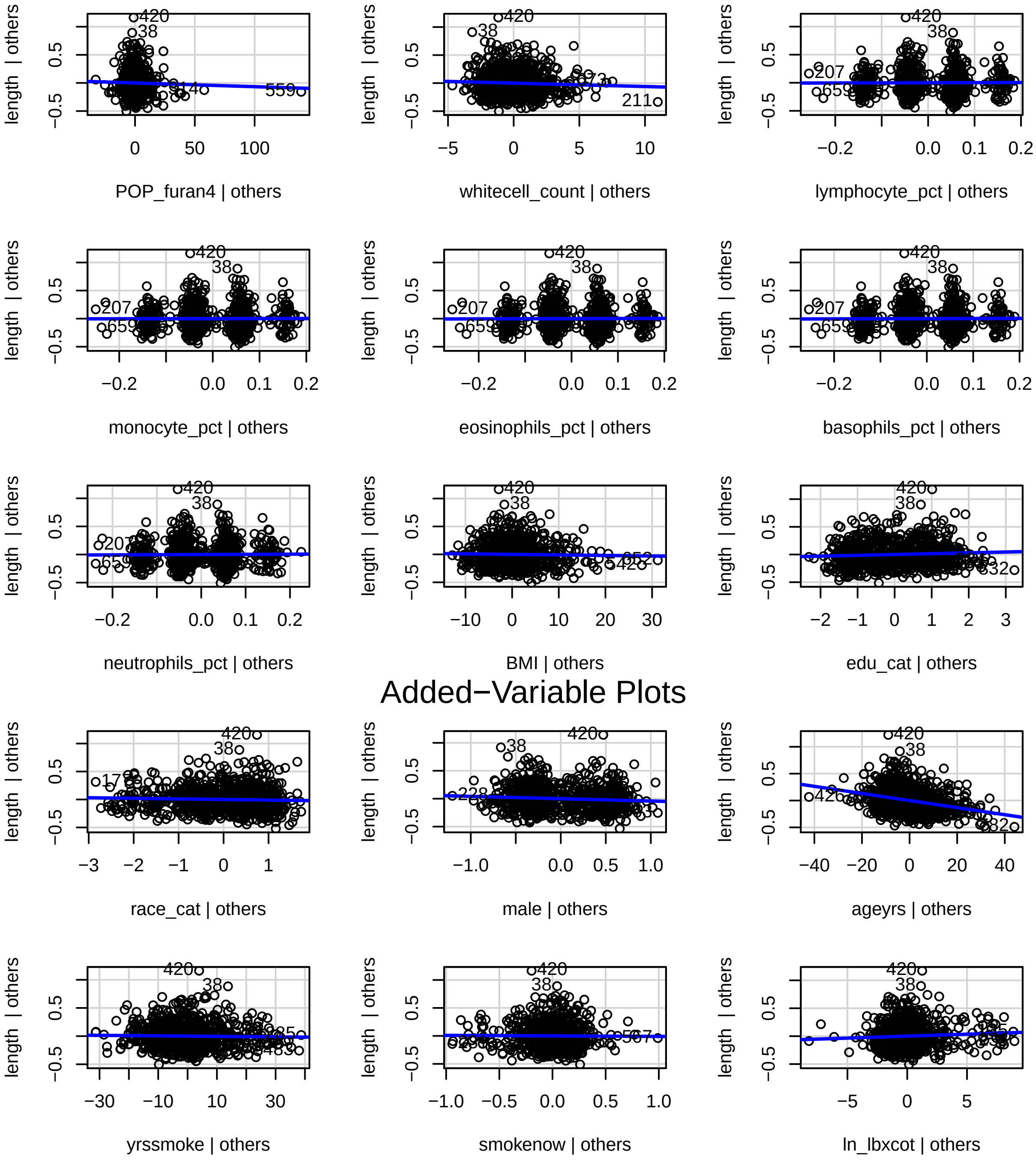


```
# Judy's work Part 2
# testing non-linearity in MLR
library(car)
```

```
## Loading required package: carData
```

```
M <- lm (length ~ ., data=pollutants)
avPlots(M)
```




Added−Variable Plots