STAT 331 Final Project

Maxine, Estella, Judy, Weiwei

04/12/2021

Contents

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

We are looking to investigate the most influential factors that contribute to the average leukocyte telomere length in a person. We would like to especially look for human-adjustable factors such as whether a person smokes or exposure to persistent organic pollutants.

3 Exploratory Data Analysis

The covariates of interest from the provided dataset are

```
names(pollutants)
```

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

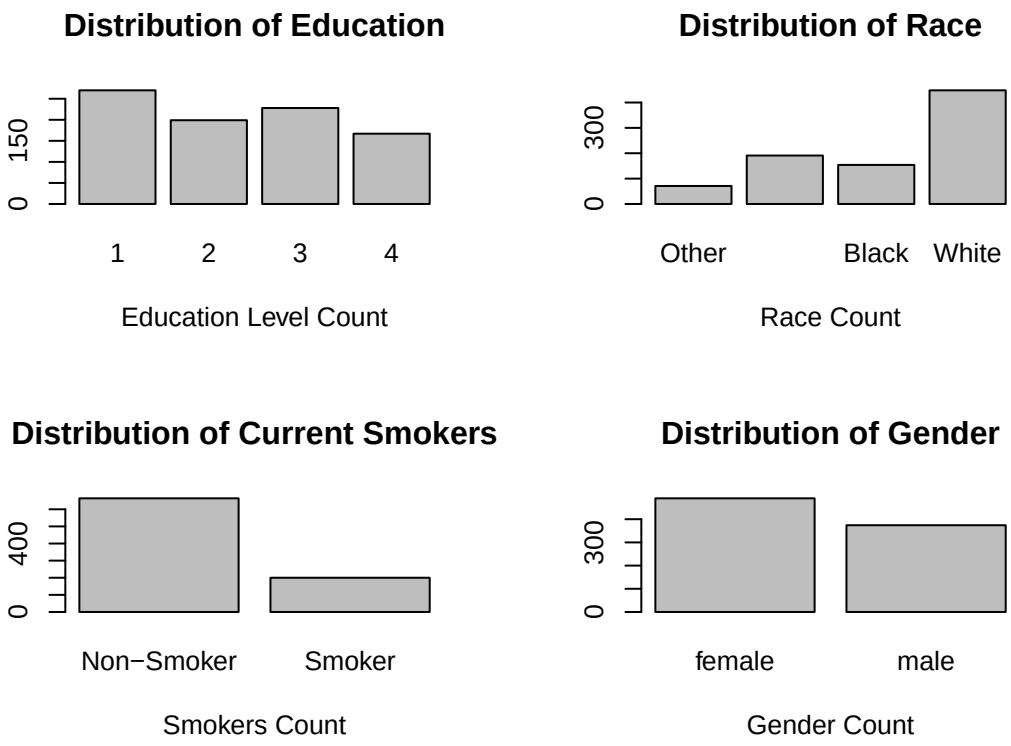
Note that “edu_cat”, “race_cat”, “male”, “smokenow” are categorical values and the rest are continuous.

3.1 Data Distribution

We shall now investigate the distribution of covariates from the supplied data.

From the output of summary statistics on the covariates (see in appendix 7.1), we observed that all values are non-negative and there are more observations with values close to 0 than values with large magnitude across all covariates.

Now we shall have a closer look at the distribution of individual covariate. For categorical data,



We may observe from the bar graphs that there are more data about non-smokers than smokers and white people than other races. There are more entries for lower education than higher, and more female than male. However, the distribution of gender and education is relatively close.

Now for continuous data, we made boxplots to see the distribution of these covariates, the plots can be found in the appendix 7.2. From these plots, we notice some extreme outliers in some concentration values of PCBs, Dioxins, and Furan. The maximum values are sometimes over double the magnitude of the second largest.

However, with a little investigation in the appendix 7.3, we see that the extreme outliers across different types of PCB mostly came from one observation.

```
pollutants[436, 3:12]
```

```
##     POP_PCB2 POP_PCB3 POP_PCB4 POP_PCB5 POP_PCB6 POP_PCB7 POP_PCB8 POP_PCB9
## 436   165000   123000   487000   708000   319000   127000   187000   144000
##     POP_PCB10 POP_PCB11
## 436       131       137
```

This observation contributes to the maximum value for PCB1 to PCB6, as well as PCB8 and PCB9

Similarly, the most extreme outliers from Dioxin and Furan also came from the same entry of data:

- Entry 285 contain the highest value for Dioxin 1 and 3, which are the two extreme outliers as we can see from the boxplots
- Entry 559 contain the highest value for Furan 2 and 4, where Furan 4 has an extreme outlier

Other covariates, as we see from the boxplots, do not have outliers that are as extreme as those from pollutant data. We further observe that they do not have a common entry that contributes to the outliers.

3.2 Multicolinearity

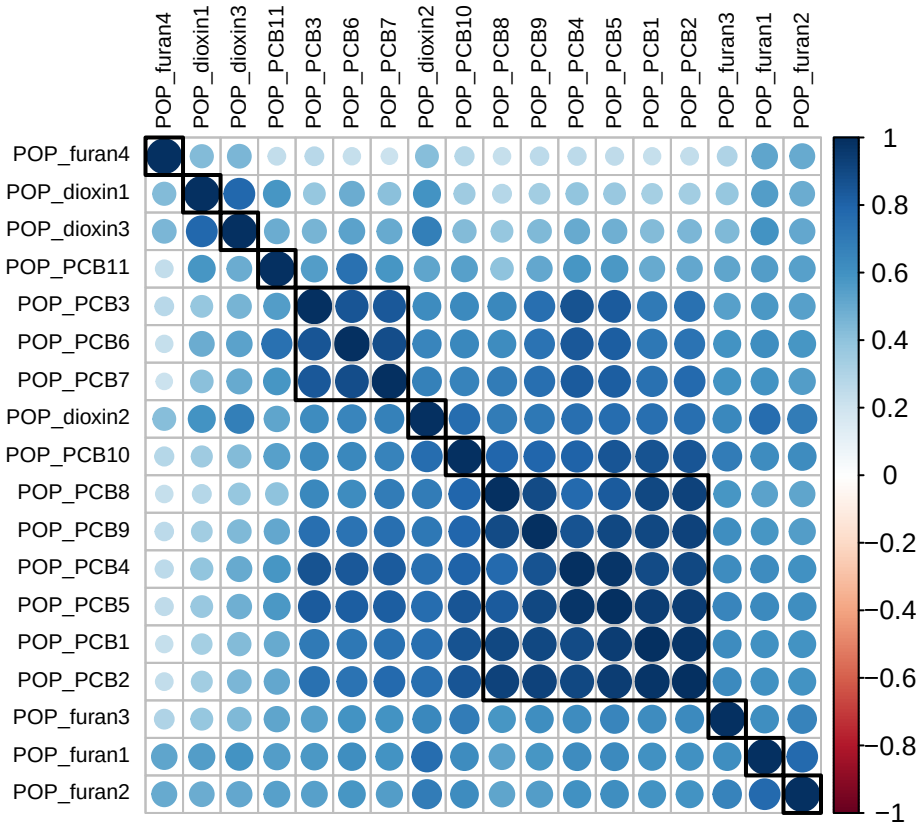
We learned that severe multicollinearity between covariates could result in unstable and sensitive coefficient estimates. Therefore, in this section, we will investigate correlations among values that we may expect multi-

collinearity to appear, such as between different types of organic pollutants POP_PCB1–11, POP_dioxin1-3, Pop_furan1–4, as well as white blood cell components.

To obtain the heatmaps that visualize correlations among covariates, we first computed Spearman correlations for each pair of covariates of interest and represented the measured values through gradients of a color scheme. In our example, blue refers to positive correlations and red, negative. Furthermore, the darker colours signify a higher correlation among the covariates. Finally, we clustered variables with higher correlations together such that the covariates within the same rectangles are highly correlated such that they may have dependencies on each other.

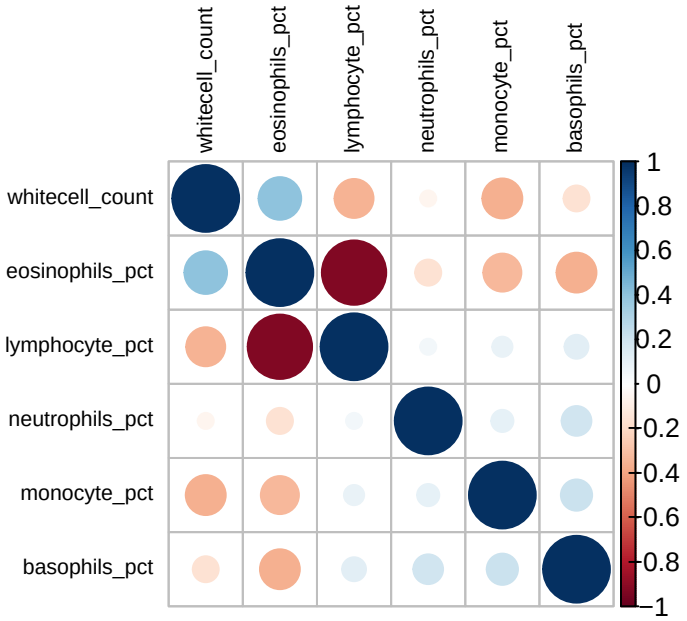
3.2.1 Correlation among Persistent Pollution

```
## corrplot 0.84 loaded
```



Based on the above plot, we noticed the correlations mostly exist among the organic pollutants of the same kind. Specifically, the correlations among POP_PCB3,6,7 and POP_PCB8,9,4,5,1,2 are higher than others.

3.2.2 Correlation between White Blood Cells



From the graph above, we see that there is no strong positive correlation among the components of white blood cells, however, there is a strong negative correlation between lymphocytes and eosinophils percentage in the given data.

To further investigate how these correlations affect the observed data, we shall consider adding interaction terms to our model, and performed p-test to check their statistical significance.

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.

```
train_data <- pollutants[1:600,]
test_data <- pollutants[601:nrow(pollutants),]
```

Linearity is one of the four assumptions in a linear regression model. To assess whether any covariate has a nonlinearity relationship with the outcome in the multiple linear regression model, added-variable plots(avPlot) are used, as shown in 7.4. The plots isolate the relationship between the outcome and each of the covariates after adjusting for the other covariate. If the plot of the outcome versus a covariate x has a nonlinear shape, the idealist regression model should include a higher power of x, for example x^2, besides the main effect. In this case, as shown in the avPlots, since all plots have a linear shape, the outcome is expected to have a linear relationship with all of the covariates. Therefore, the model constructed in this report does not consider nonlinear term for any of the covariates.

As the given dataset is relatively large, we may assume the data is approximately Normally distributed due to the Central Limit Theorem.

```
# Estella's work 3
f <- as.formula(
  paste("length", paste("(", paste(POP_chemicals, collapse = "+"), ")^2"), sep="~"))
m_chem <- lm(f, data = pollutants)
# summary(m_chem)
```

```r
# Estella's work 4
# setting threshold of pvalue to be 0.05 and assess possible interaction terms
pvalues <- summary(m_chem)$coefficients[19:nrow(summary(m_chem)$coefficients),4]
p_threshold = 0.05
selected <- which(pvalues <= p_threshold)
names(selected)
```

```
##  [1] "POP_PCB1:POP_PCB9"    "POP_PCB2:POP_PCB6"    "POP_PCB2:POP_PCB8"
##  [4] "POP_PCB2:POP_PCB9"    "POP_PCB2:POP_PCB10"   "POP_PCB2:POP_PCB11"
##  [7] "POP_PCB2:POP_furan4"  "POP_PCB3:POP_furan3"  "POP_PCB4:POP_PCB9"
## [10] "POP_PCB4:POP_dioxin3" "POP_PCB5:POP_PCB11"   "POP_PCB5:POP_dioxin3"
## [13] "POP_PCB6:POP_PCB8"    "POP_PCB7:POP_PCB8"    "POP_PCB7:POP_PCB9"
```

```r
#stepwise parameters selection without any interaction terms
M0 <- lm(length ~ 1, data = train_data) # minimal model
Mfull <- lm(length ~ ., data= train_data)

## 2 corresponds to AIC
## log(n) corresponds to BIC

# stepwise AIC
Mstart <- lm(length ~ ., data= train_data)
system.time({
  MAIC <- step(object = Mstart,
               scope = list(lower = M0, upper = Mfull),
               direction = "both", trace = 0, k = 2)
})
```

```
##    user  system elapsed
##   0.926   0.101   1.117
```

```r
#stepwiseBIC
system.time({
  MBIC <- step(object = Mstart,
               scope = list(lower = M0, upper = Mfull),
               direction = "both", trace = 0, k = log(nrow(train_data)))
})
```

```
##    user  system elapsed
##   1.000   0.101   1.185
```

```r
#stepwiseB_Adjusted R2
MAIC
```

```
##
## Call:
## lm(formula = length ~ POP_PCB1 + POP_PCB10 + POP_furan1 + POP_furan2 +
##     whitecell_count + monocyte_pct + edu_cat + race_cat + male +
##     ageyrs + ln_lbxcot, data = train_data)
##
## Coefficients:
##     (Intercept)         POP_PCB1        POP_PCB10       POP_furan1
##       1.443e+00       -5.602e-07        1.780e-03       -6.532e-03
##      POP_furan2  whitecell_count     monocyte_pct         edu_cat2
##       8.968e-03       -1.029e-02       -6.643e-03        4.105e-02
##        edu_cat3         edu_cat4  race_catMexican    race_catBlack
##       6.188e-02        8.254e-02       -3.635e-03        3.584e-02
```

```
##   race_catWhite         malemale           ageyrs        ln_lbxcot
##      -4.701e-02       -4.513e-02       -5.820e-03        7.573e-03
```

```
MBIC
```

```
##
## Call:
## lm(formula = length ~ POP_furan3 + ageyrs, data = train_data)
##
## Coefficients:
## (Intercept)   POP_furan3       ageyrs
##    1.355743     0.005969    -0.006922
```

```
# stepwise parameters selection with any interaction terms
M0 <- lm(length ~ 1, data = train_data) # minimal model

# tail to remove length column
single <- paste(tail(colnames(train_data),-1), collapse = " + ")
# tail to remove intercept column
interaction <- paste(tail(names(selected),-1), collapse = " + ")
f_interaction <- as.formula(
  paste("length", paste("(", single,"+", interaction, ")"), sep = " ~"))

Mfull <- lm(f_interaction, data = train_data)
Mstart <- lm(f_interaction, data = train_data)

# stepwise AIC
Mstart <- lm(length ~ ., data= train_data)
system.time({
  MAIC_Interaction <- step(object = Mstart,
                           scope = list(lower = M0, upper = Mfull),
                           direction = "both", trace = 0, k = 2)
})
```

```
##    user  system elapsed
##   0.964   0.091   1.100
```

```
#stepwiseBIC
system.time({
  MBIC_Interaction <- step(object = Mstart,
                           scope = list(lower = M0, upper = Mfull),
                           direction = "both", trace = 0,
                           k = log(nrow(train_data)))
})
```

```
##    user  system elapsed
##   0.965   0.082   1.083
```

```
#stepwiseB_Adjusted R2
MAIC_Interaction
```

```
##
## Call:
## lm(formula = length ~ POP_PCB1 + POP_PCB10 + POP_furan1 + POP_furan2 +
##     whitecell_count + monocyte_pct + edu_cat + race_cat + male +
##     ageyrs + ln_lbxcot, data = train_data)
##
## Coefficients:
```

```
##     (Intercept)         POP_PCB1        POP_PCB10       POP_furan1
##       1.443e+00       -5.602e-07        1.780e-03       -6.532e-03
##      POP_furan2  whitecell_count     monocyte_pct         edu_cat2
##       8.968e-03       -1.029e-02       -6.643e-03        4.105e-02
##        edu_cat3         edu_cat4  race_catMexican    race_catBlack
##       6.188e-02        8.254e-02       -3.635e-03        3.584e-02
##   race_catWhite         malemale           ageyrs        ln_lbxcot
##      -4.701e-02       -4.513e-02       -5.820e-03        7.573e-03
```

```
MBIC_Interaction
```

```
##
## Call:
## lm(formula = length ~ POP_furan3 + ageyrs, data = train_data)
##
## Coefficients:
## (Intercept)   POP_furan3       ageyrs
##    1.355743     0.005969    -0.006922
```

```
# mxn's work
predAIC <- predict(MAIC, newdata=test_data)
predBIC <- predict(MBIC, newdata=test_data)
predAICInteraction <- predict(MAIC_Interaction, newdata=test_data)
predBICInteraction <- predict(MBIC_Interaction, newdata=test_data)

mean((test_data$length - predAIC)^2)
```

```
## [1] 0.05336494
```

```
mean((test_data$length - predBIC)^2)
```

```
## [1] 0.04804827
```

```
mean((test_data$length - predAICInteraction)^2)
```

```
## [1] 0.05336494
```

```
mean((test_data$length - predBICInteraction)^2)
```

```
## [1] 0.04804827
```

5 Results

Report on the findings of your analysis

6 Discussion

Comment on your findings/conclusions; describe any limitations of your analysis.

We have considered the multicollinearity and interactions within the eleven PCB covariates and similarly for the three dioxin covariates and four furan covariates. However, the multicollinearity and interactions between these eighteen exposure covariates and other covariates are not considered. It is expected that there does not exist any causal relationship between exposure covariates and other covariates since the former relates to the surrounding environment and the latter relates to personal characteristics. For example, it's believed that the concentration of POP_PCB10 is unrelated to the value of ageyrs and BMI.

Besides, the report has analyzed whether the outcome has a nonlinear relationship with any of the covariates. However, it has not fully analyzed whether any of the remaining assumptions for linear regression models are broken, which include independence, normality and homoscedasticity. These assumptions can be further tested with plots and investigations of the residuals.

7 Appendix

7.1 Data Summary

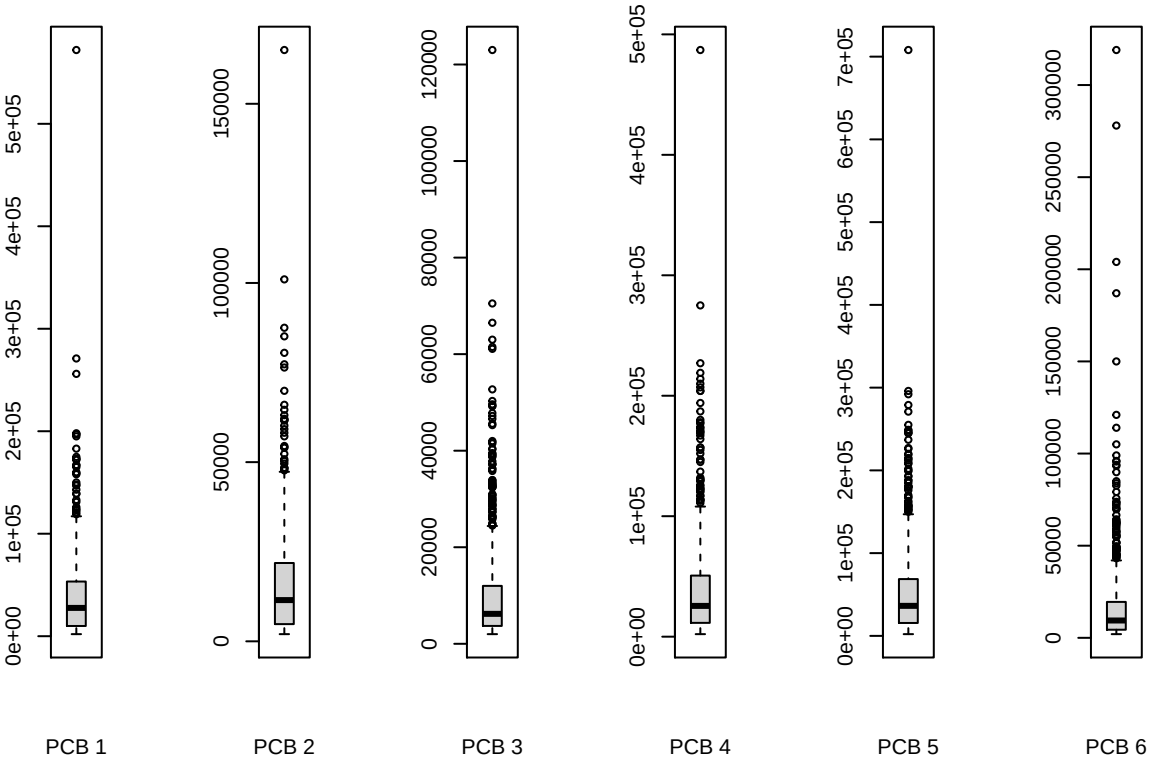
Looking at the useful metrics for the data

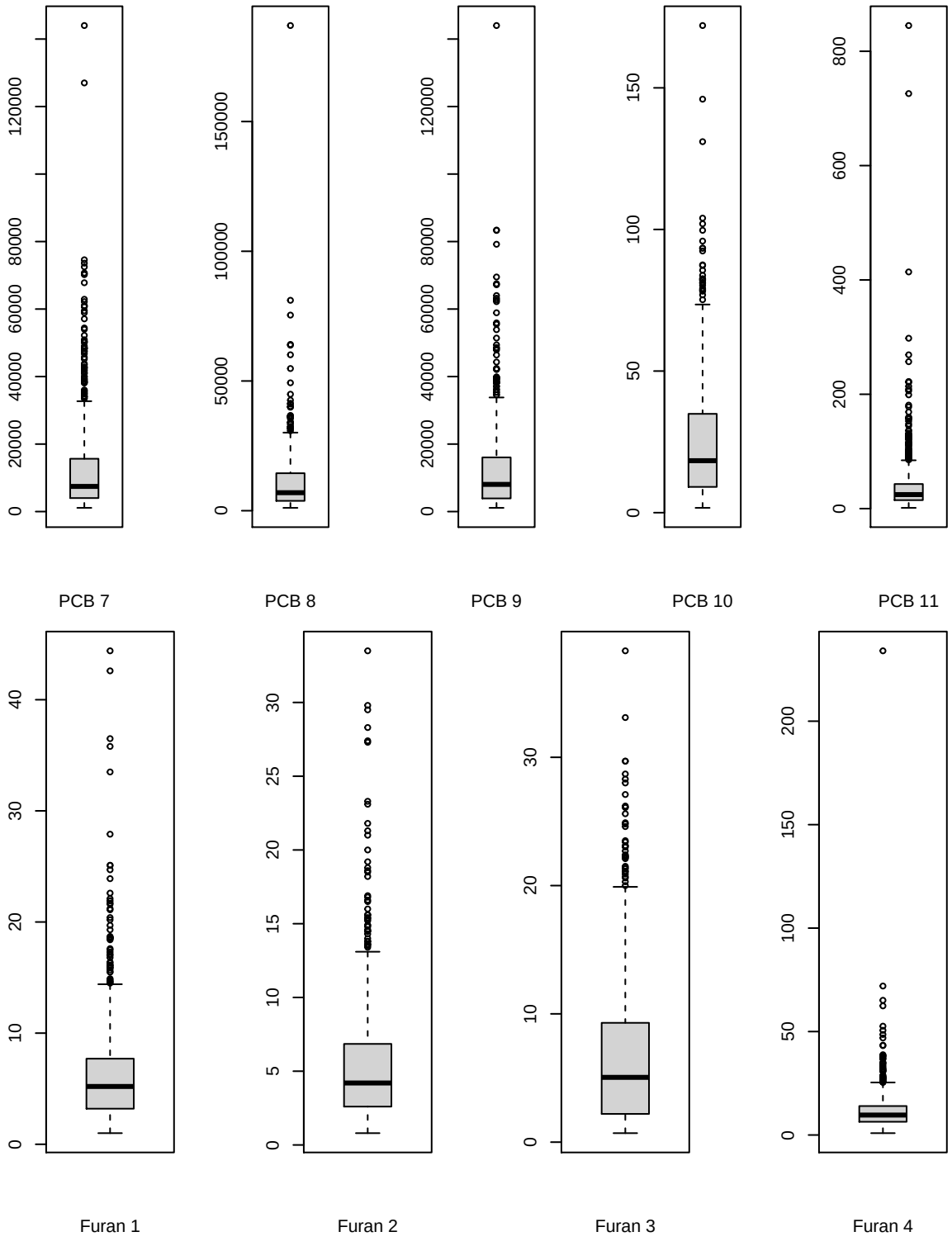
```
summary(pollutants)
```

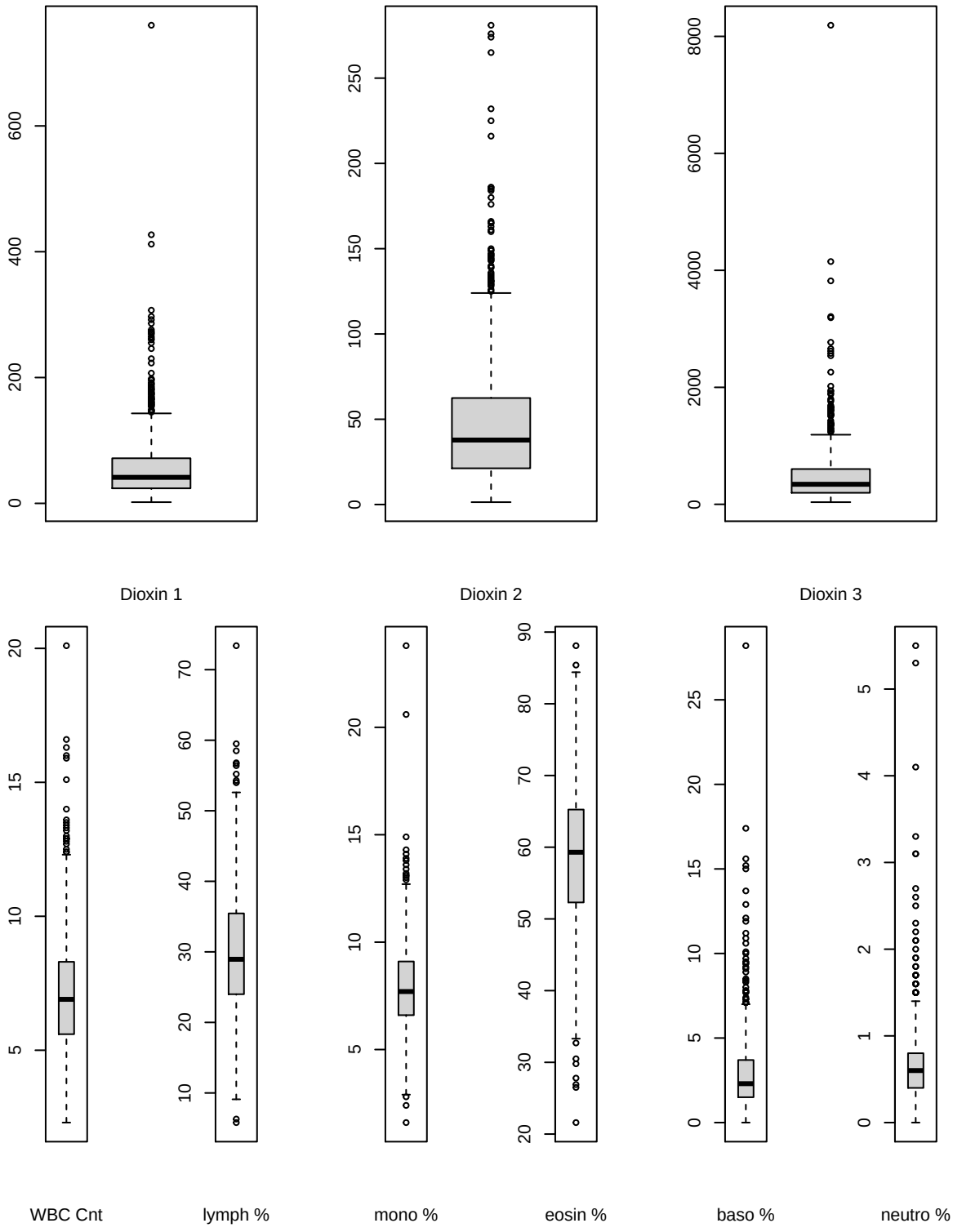
```
##      length          POP_PCB1         POP_PCB2         POP_PCB3
##  Min.   :0.5266   Min.   :  2000   Min.   :  2000   Min.   :  2000
##  1st Qu.:0.8754   1st Qu.:  9975   1st Qu.:  4800   1st Qu.:  3700
##  Median :1.0286   Median : 27600   Median : 11500   Median :  6200
##  Mean   :1.0543   Mean   : 38082   Mean   : 15637   Mean   : 10158
##  3rd Qu.:1.2095   3rd Qu.: 53325   3rd Qu.: 21825   3rd Qu.: 12000
##  Max.   :2.3512   Max.   :572000   Max.   :165000   Max.   :123000
##     POP_PCB4         POP_PCB5         POP_PCB6         POP_PCB7
##  Min.   :  2100   Min.   :  2100   Min.   :  2000   Min.   :  1100
##  1st Qu.: 11475   1st Qu.: 15600   1st Qu.:  4400   1st Qu.:  4000
##  Median : 25550   Median : 36300   Median :  9400   Median :  7450
##  Mean   : 38456   Mean   : 52650   Mean   : 16820   Mean   : 12682
##  3rd Qu.: 50650   3rd Qu.: 68625   3rd Qu.: 19500   3rd Qu.: 15625
##  Max.   :487000   Max.   :708000   Max.   :319000   Max.   :144000
##     POP_PCB8         POP_PCB9        POP_PCB10        POP_PCB11
##  Min.   :  1100   Min.   :  1100   Min.   :  1.70   Min.   :  1.30
##  1st Qu.:  3800   1st Qu.:  3900   1st Qu.:  9.10   1st Qu.: 14.80
##  Median :  6950   Median :  8050   Median : 18.35   Median : 24.50
##  Mean   : 10530   Mean   : 12220   Mean   : 24.49   Mean   : 38.15
##  3rd Qu.: 14425   3rd Qu.: 16025   3rd Qu.: 34.90   3rd Qu.: 42.95
##  Max.   :187000   Max.   :144000   Max.   :172.00   Max.   :845.00
##   POP_dioxin1      POP_dioxin2      POP_dioxin3       POP_furan1
##  Min.   :  1.90   Min.   :  1.40   Min.   :  36.8   Min.   : 1.000
##  1st Qu.: 23.90   1st Qu.: 21.27   1st Qu.: 197.0   1st Qu.: 3.200
##  Median : 41.35   Median : 37.80   Median : 342.5   Median : 5.200
##  Mean   : 57.65   Mean   : 47.81   Mean   : 494.4   Mean   : 6.371
##  3rd Qu.: 71.62   3rd Qu.: 62.42   3rd Qu.: 603.0   3rd Qu.: 7.700
##  Max.   :760.00   Max.   :281.00   Max.   :8190.0   Max.   :44.400
##    POP_furan2       POP_furan3       POP_furan4     whitecell_count
##  Min.   : 0.800   Min.   : 0.700   Min.   :  0.90   Min.   : 2.300
##  1st Qu.: 2.600   1st Qu.: 2.200   1st Qu.:  6.40   1st Qu.: 5.600
##  Median : 4.200   Median : 5.050   Median :  9.65   Median : 6.900
##  Mean   : 5.390   Mean   : 6.669   Mean   : 11.54   Mean   : 7.191
##  3rd Qu.: 6.825   3rd Qu.: 9.300   3rd Qu.: 14.00   3rd Qu.: 8.300
##  Max.   :33.500   Max.   :38.300   Max.   :234.00   Max.   :20.100
##  lymphocyte_pct   monocyte_pct    eosinophils_pct basophils_pct
##  Min.   : 5.80   Min.   : 1.600   Min.   :21.60   Min.   : 0.000
##  1st Qu.:24.00   1st Qu.: 6.600   1st Qu.:52.35   1st Qu.: 1.500
##  Median :28.95   Median : 7.700   Median :59.30   Median : 2.300
##  Mean   :29.92   Mean   : 7.936   Mean   :58.62   Mean   : 2.903
##  3rd Qu.:35.42   3rd Qu.: 9.100   3rd Qu.:65.22   3rd Qu.: 3.700
##  Max.   :73.40   Max.   :23.800   Max.   :88.10   Max.   :28.200
##  neutrophils_pct       BMI        edu_cat    race_cat       male
##  Min.   :0.0000   Min.   :16.16   1:270   Other  : 71   female:490
##  1st Qu.:0.4000   1st Qu.:23.88   2:199   Mexican:191   male  :374
##  Median :0.6000   Median :27.38   3:228   Black  :154
```

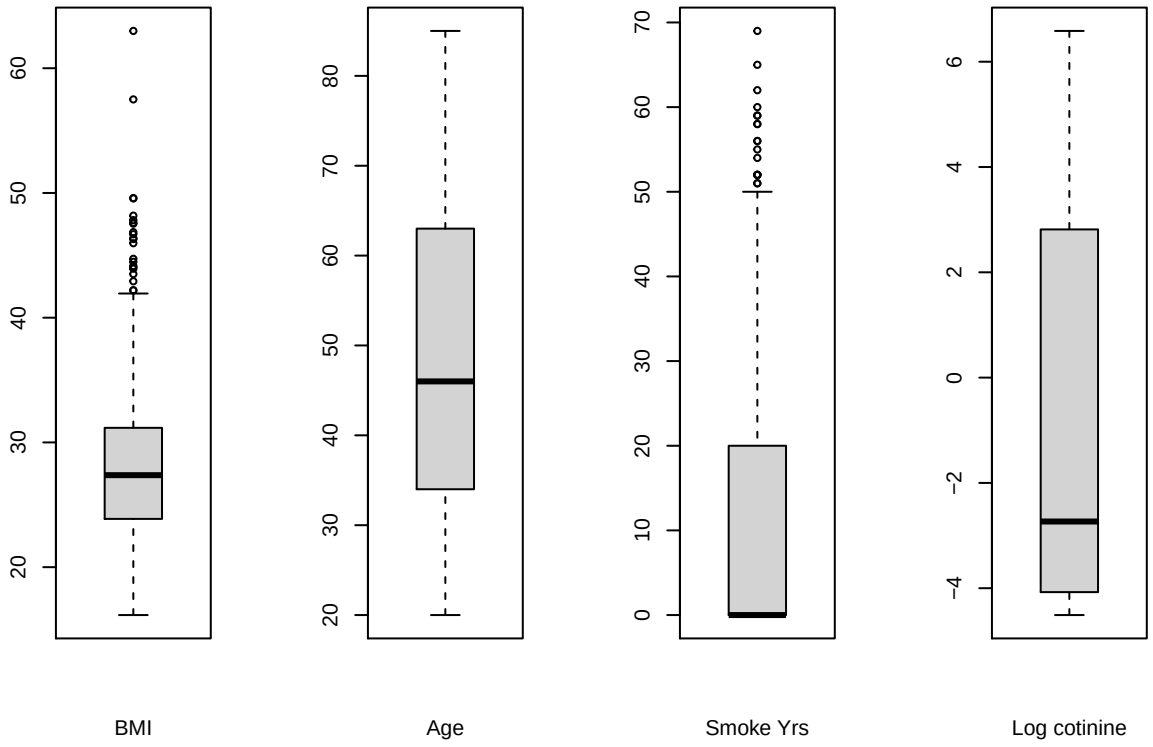
```
##  Mean   :0.6669   Mean   :28.09   4:167   White  :448
##  3rd Qu.:0.8000   3rd Qu.:31.17
##  Max.   :5.5000   Max.   :62.99
##      ageyrs         yrssmoke          smokenow     ln_lbxcot
##  Min.   :20.00   Min.   : 0.0   Non-Smoker:664   Min.   :-4.5099
##  1st Qu.:34.00   1st Qu.: 0.0   Smoker    :200   1st Qu.:-4.0745
##  Median :46.00   Median : 0.0                    Median :-2.7334
##  Mean   :48.36   Mean   :10.6                    Mean   :-0.9804
##  3rd Qu.:63.00   3rd Qu.:20.0                    3rd Qu.: 2.8000
##  Max.   :85.00   Max.   :69.0                    Max.   : 6.5848
```

7.2 Boxplots




0
20000
40000
60000
80000
120000
PCB 7
0
50000
100000
150000
PCB 8
0
20000
40000
60000
80000
120000
PCB 9
0
50
100
150
PCB 10
0
200
400
600
800
PCB 11
0
10
20
30
40
Furan 1
0
5
10
15
20
25
30
Furan 2
0
10
20
30
Furan 3
0
50
100
150
200
Furan 4


Dioxin 1
Dioxin 2
Dioxin 3
WBC Cnt
lymph %
mono %
eosin %
baso %
neutro %



7.3 Outlier Entries

Here we will find entries where outliers for different covariate occurred.

```
pollutant_mat = data.matrix(pollutants, rownames.force = NA)

max_PCB_idx = c()
for (c in 2:12) {
  max_PCB_idx[c-1] = which.max(pollutant_mat[, c])
}
max_PCB_idx
```

```
##  [1] 436 436 436 436 436 436 426 436 436 298 272
```

```
max_dioxin_idx = c()
for (c in 13:15) {
  max_dioxin_idx[c-12] = which.max(pollutant_mat[, c])
}
max_dioxin_idx
```

```
## [1] 285 573 285
```

```
max_furan_idx = c()
for (c in 16:19) {
  max_furan_idx[c-15] = which.max(pollutant_mat[, c])
}
max_furan_idx
```

```
## [1] 230 559 590 559
```

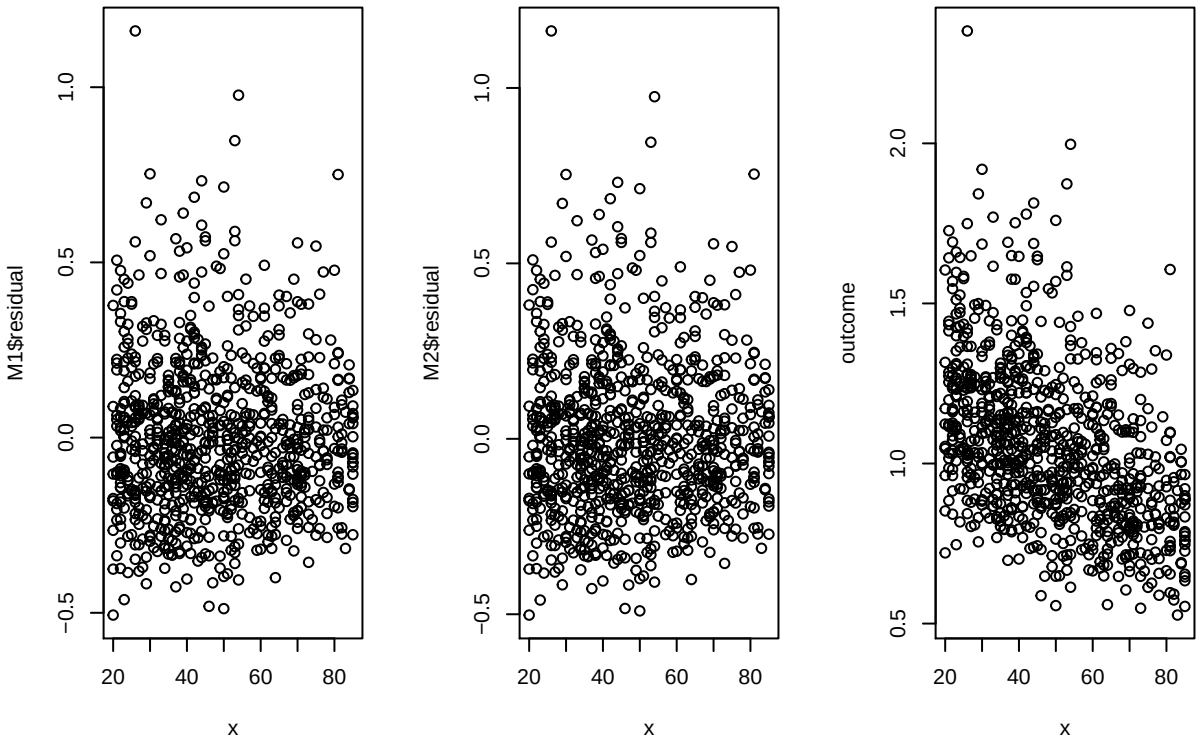
```r
max_WBC_idx = c()
for (c in 20:25) {
  max_WBC_idx[c-19] = which.max(pollutant_mat[, c])
}
max_WBC_idx
```

```
## [1] 211 766 440 782 739 415
```

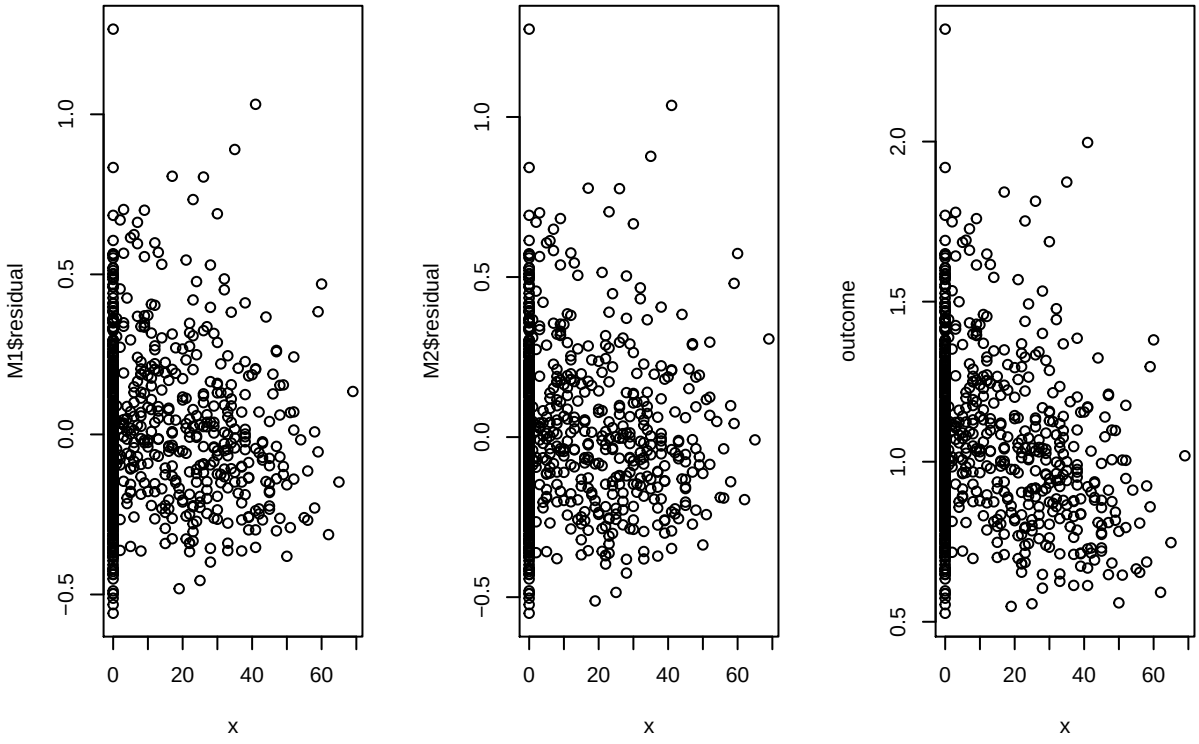
7.4 AvPlots

```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)
}
```

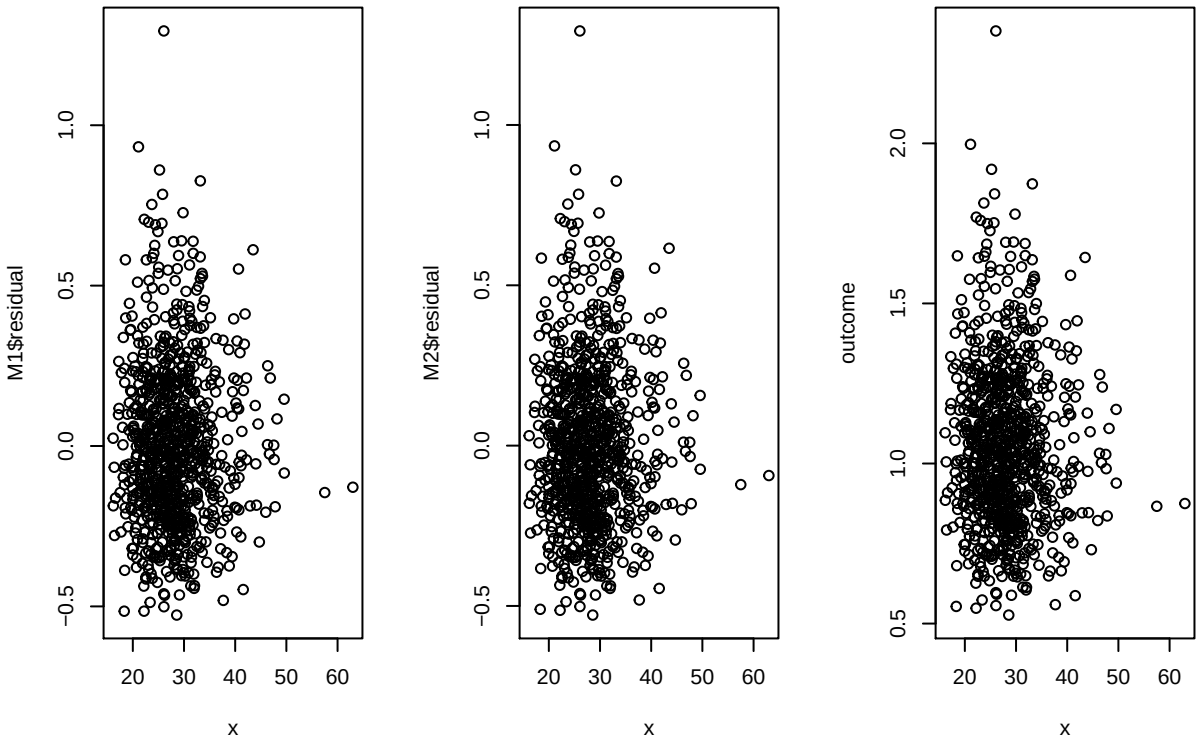
```
## [1] "residual for M1:  0.224172364185412"
## [1] "residual for M2:  0.22429269961392"
```



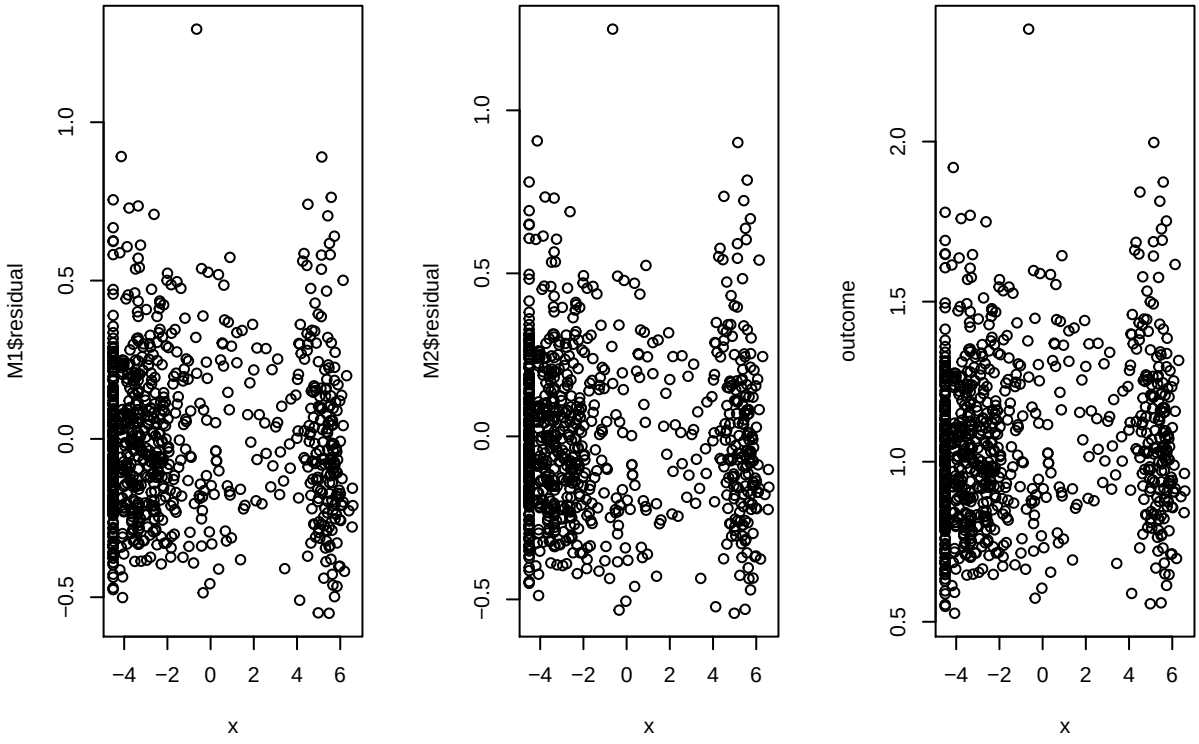

```
## [1] "residual for M1:  0.246320733146214"
## [1] "residual for M2:  0.245622720856213"
```



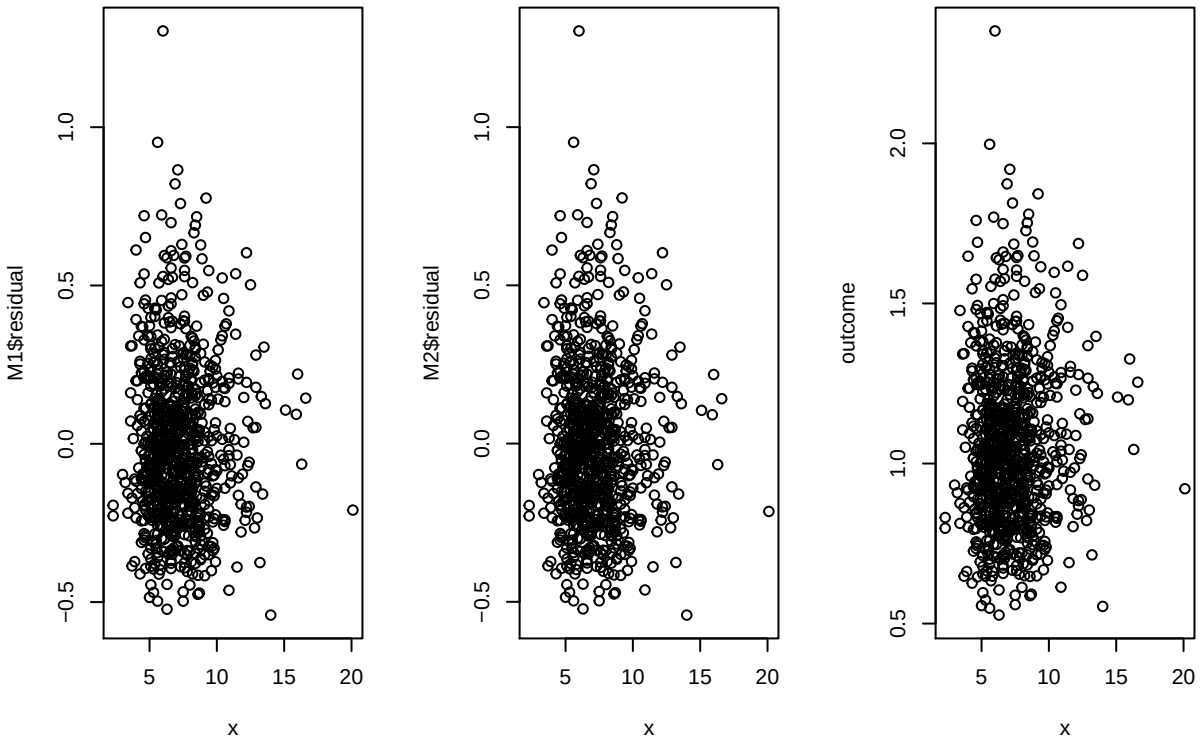

```
## [1] "residual for M1:  0.250228706427173"
## [1] "residual for M2:  0.25036248052387"
```



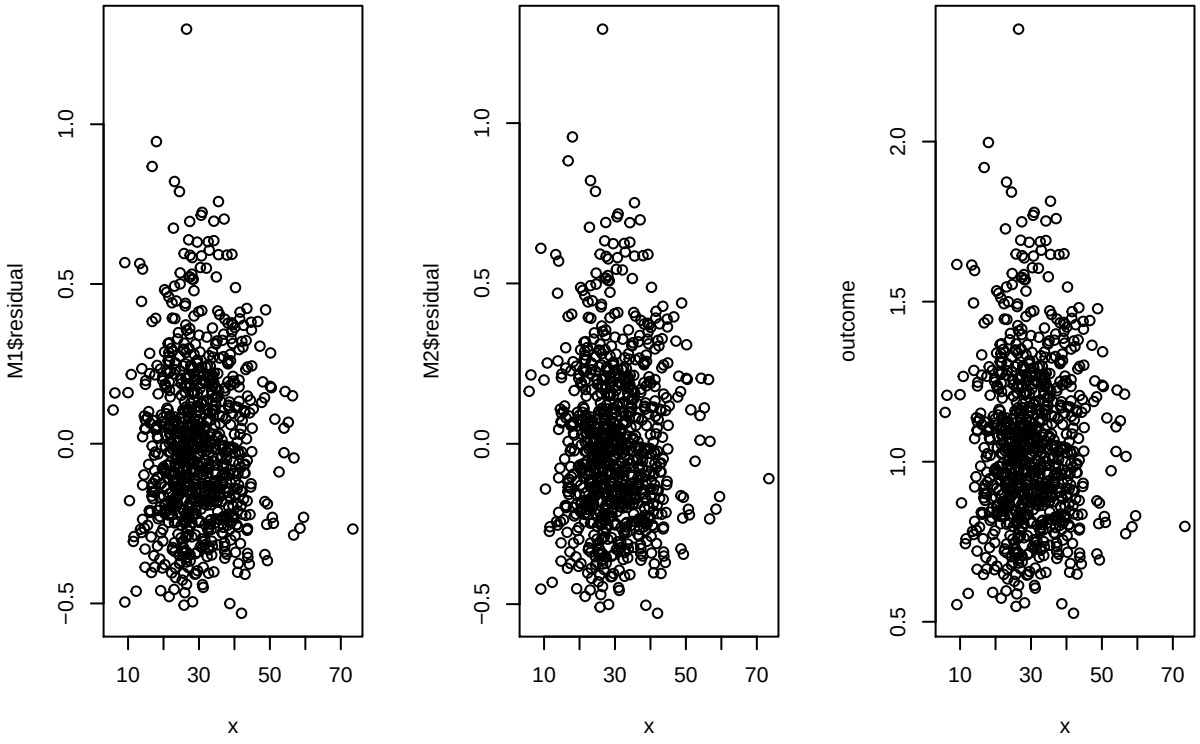

```
## [1] "residual for M1:  0.248212063673837"
## [1] "residual for M2:  0.24710732733351"
```



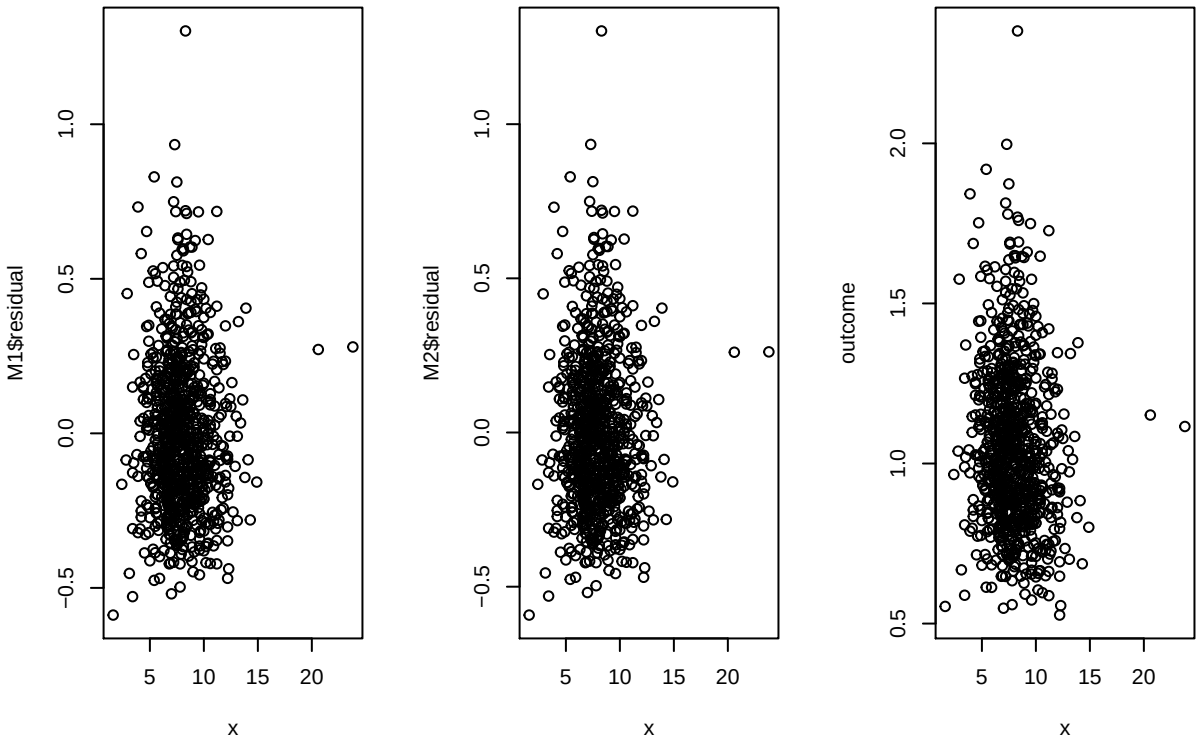

```
## [1] "residual for M1:  0.250065445847753"
## [1] "residual for M2:  0.250210403543218"
```



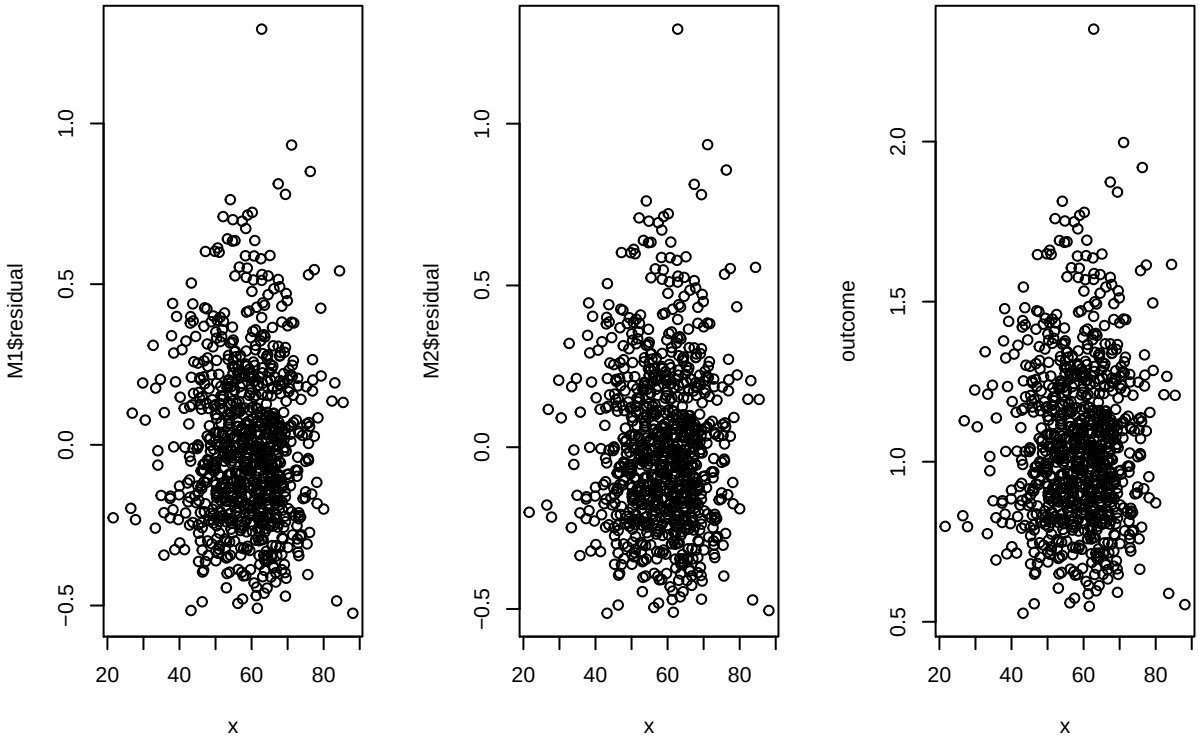

```
## [1] "residual for M1:  0.250373616826691"
## [1] "residual for M2:  0.250255208638358"
```



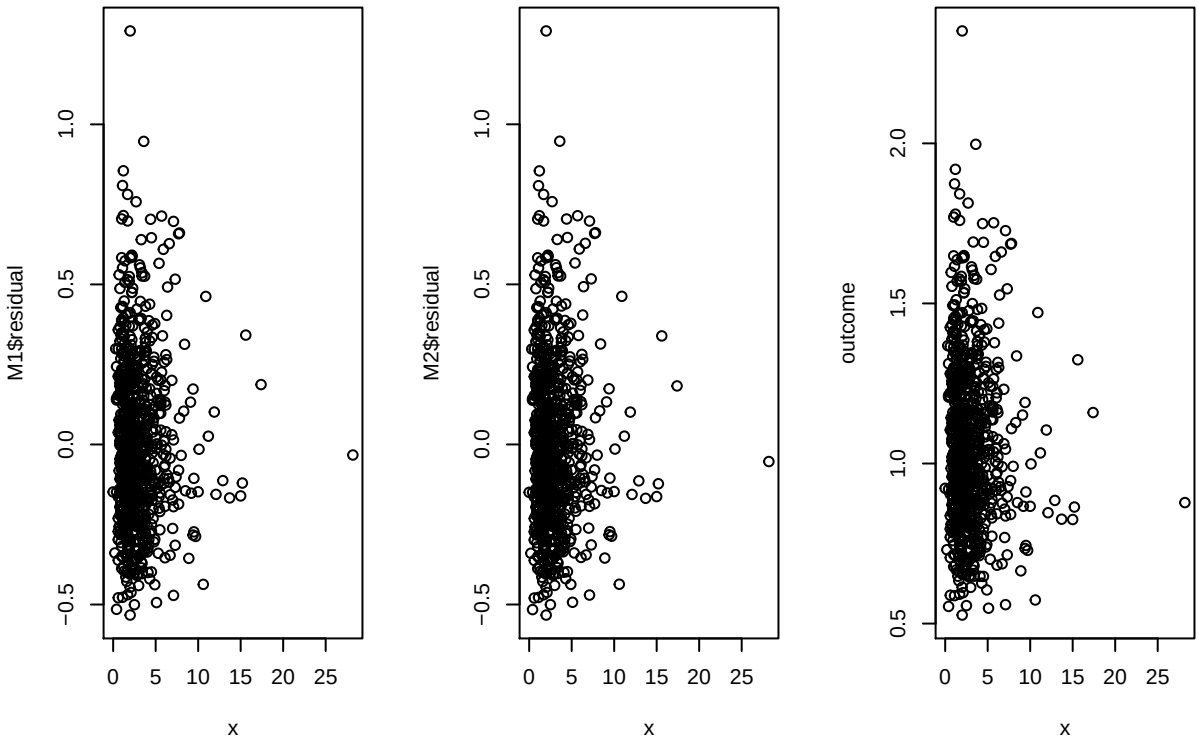

```
## [1] "residual for M1:  0.248704466454944"
## [1] "residual for M2:  0.248847192837983"
```



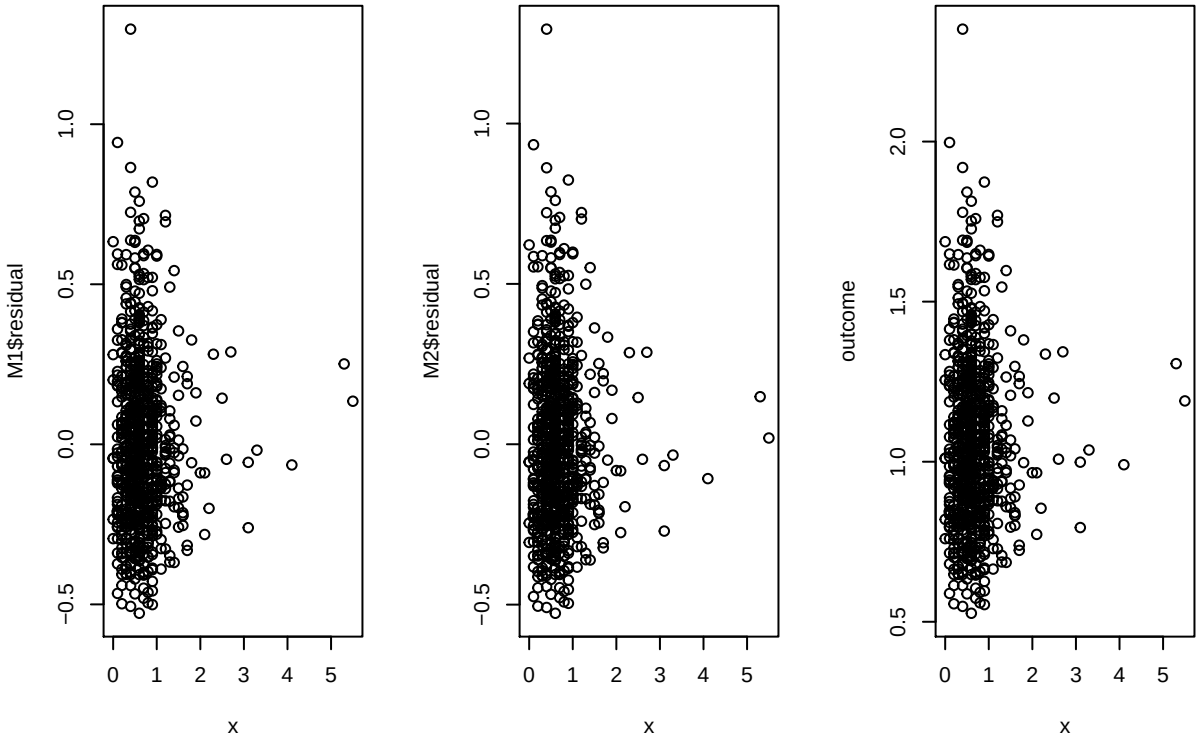

```
## [1] "residual for M1:  0.25026710930793"
## [1] "residual for M2:  0.250393729526099"
```




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



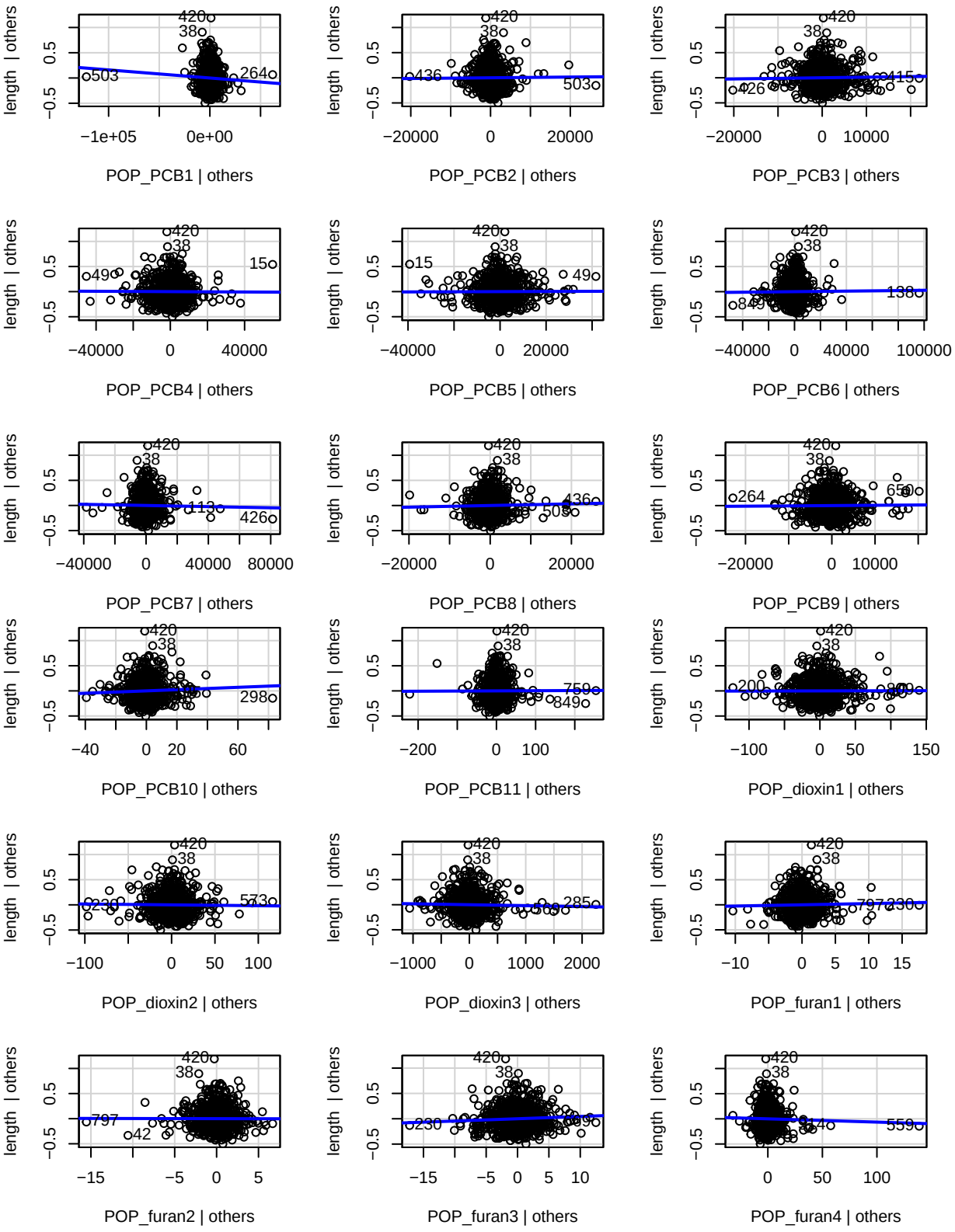

```
## [1] "residual for M1:  0.250382476371691"
## [1] "residual for M2:  0.25042580861039"
```

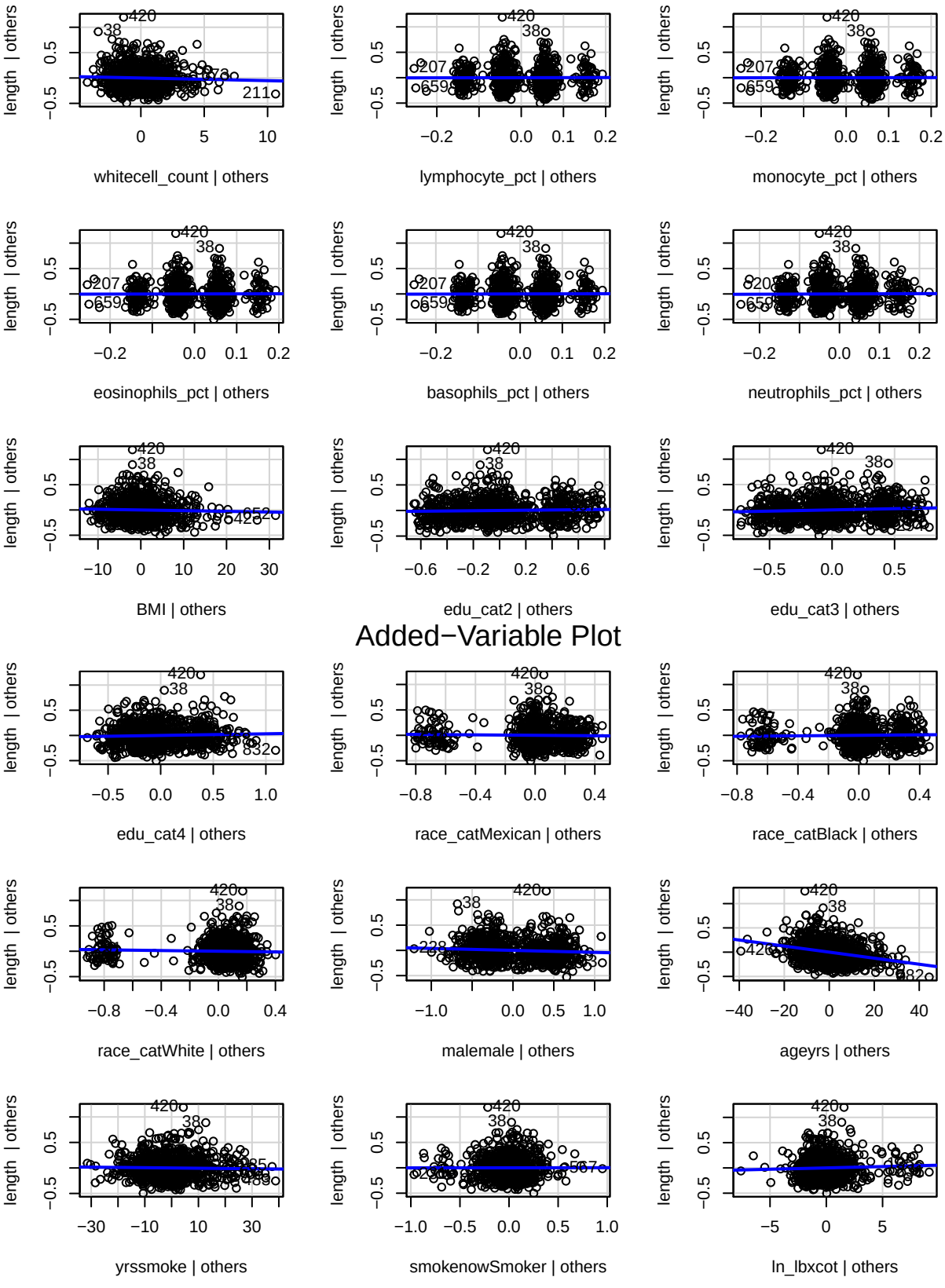



```
# Judy's work Part 2
# testing non-linearity in MLR
library(car)
```

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

```
M <- lm (length ~ ., data=pollutants)
avPlots(M, main="Added-Variable Plot")
```




length | others
whitecell_count | others
lymphocyte_pct | others
monocyte_pct | others
eosinophils_pct | others
basophils_pct | others
neutrophils_pct | others
BMI | others
edu_cat2 | others
edu_cat3 | others
Added−Variable Plot
edu_cat4 | others
race_catMexican | others
race_catBlack | others
race_catWhite | others
malemale | others
ageyrs | others
yrssmoke | others
smokenowSmoker | others
ln_lbxcot | others