STAT 331 Final Project

Statistical Modeling and Analysis Results for pollution.Rdata
Submitted to: Prof. Glen McGee

Report Prepared By: Group 43
Qin, Ziqi: 20758068
Wu, Xuanxiao: 20702042
Yang, Yun Fan: 20841752
Zhu, yinyao: 20726577

August 4, 2021

Summary

In this project, we have a total of 1000 samples and 80 variables. Based on these samples, we found that some variables had no effect on birthweight, while others were strongly associated with birthweight. In order to achieve our goal of detecting the variables that have the significant impact on birthweight, we built several candidate models. Finally, we found the stepwise BIC Model to be our final model as it was the most appropriate fitted model for the data. In this model, we find that *Gestational age at birth (week)*, *Child sex (female / male)*, *Perfluorooctanoate (PFOA) in mother* are very important. In addition, we found that *Gestational age at birth* had the greatest effect on birthweight, and *Triclosan (TRCS)* in mother adjusted for creatinine had the least effect, and even removing it did not actually affect birthweight.Lastly, we also looked at the chemical and non-chemical exposure separately, and found two smaller models for birthweight.

Objective

This project is about the effect of a number of variables on birthweight. The research is based on two file, the first file "codebook.csv" contains the outcome (birthweight) and 80 explanatory variables. The explanatory variables could be divided into two categories, chemical exposure measured in mother's blood/urine/hair, and outdoor exposure measured in the surrounding environment. Furthermore, the data also contains seven covariates which are maternal age, education, BMI, weight gain during pregnancy, child's year of birth, gestational age at birth, and sex. The second file is "pollution.Rdata", and it contains 1000 samples that were taken from pregnant women.

With the development of science and technology, living conditions are getting better, and life is more convenient. But in fact, the pollutants in life are also increasing, and the harmful substances that mothers can be exposed to during pregnancy are also increasing through diet, living environment, smog and so on. So we chose to use the birthweight to measure the health of the baby, to help us learn the effects of these chemicals on the baby's health. Our goal was to use the fitted model to help us find out which variable had the greatest impact on birthweight. In addition, we also wanted to know how chemicals and non-chemicals variables affect the birthweight differently.

Exploratory Data analysis

Before finding the fitting model, we first summarized the model assuming that it depends on all 80 variables, which is the null hypothesis. Then we got the results of the summary in the following picture.

```
Residual standard error: 397 on 902 degrees of freedom
Multiple R-squared:  0.4513,    Adjusted R-squared:  0.3923
F-statistic: 7.648 on 97 and 902 DF,  p-value: < 2.2e-16
```

From the picture above, we found that for our full model, the adjusted R-squared value

is 0.3923 and the degrees of freedom is 902. We also got a p-value of 2.2e-16 which is smaller than 0.0001, so we were able to reject the null hypothesis at the 5% level.

In addition, we also got another picture below showing some numerical results.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1262.74  -247.49    -7.14   226.28  1325.24
```

From the picture above, we found that the range is 2587.98 and the median is -7.14.

We take a look at the influence of these variables, *Gestational age at birth* and *Walking and/or cycling acitivity during pregnancy (frequency)* had the greatest effect on birthweight.

Gestational age at birth has a coefficient of 159.1623937.

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.591624e+02 8.155516e+00 1.951592e+01 4.947493e-71
```

Walking and/or cycling acitivity during pregnancy (frequency) has a coefficient of 159.5381784.

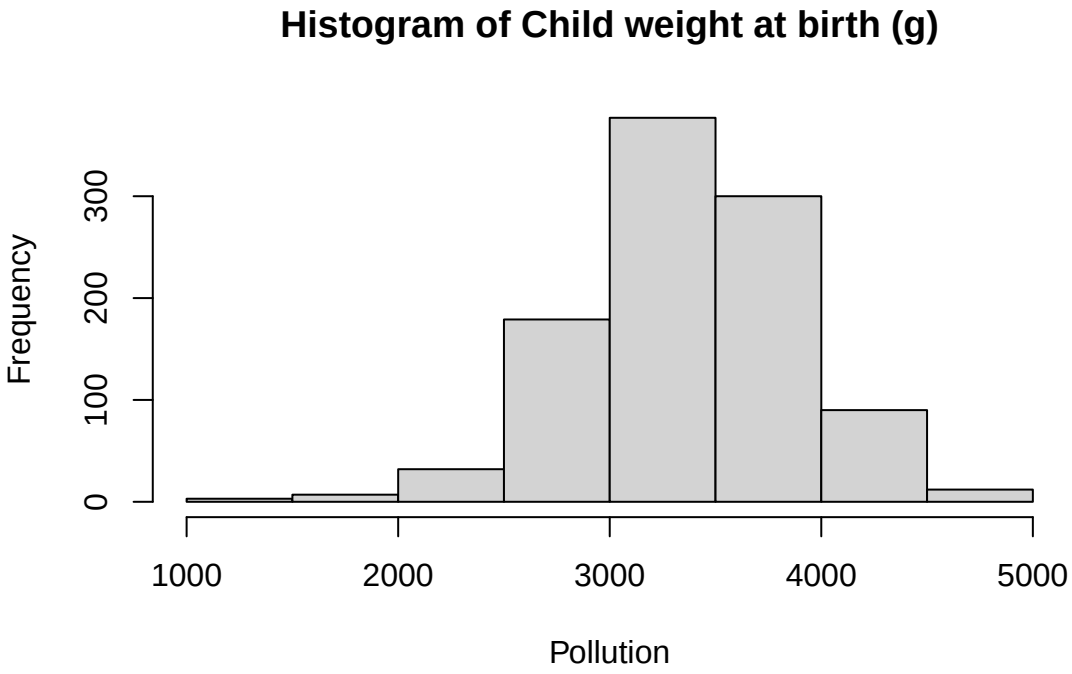
```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 159.53817842  76.99124166   2.07216009   0.03853417
```

Triclosan (TRCS) in mother adjusted for creatinine had the least effect which has a coefficient of 0.1187133.

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.11871328  4.10098432 -0.02894751  0.97691286
```

Next we checked the birthweights through the five number summary and histogram.

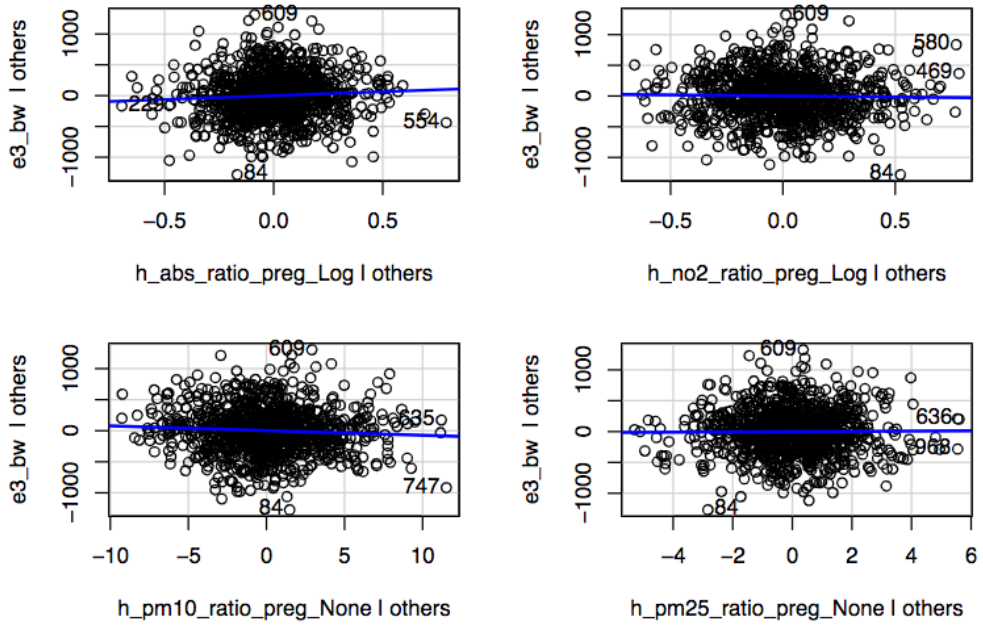
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1280    3050    3390    3378    3720    4850
```

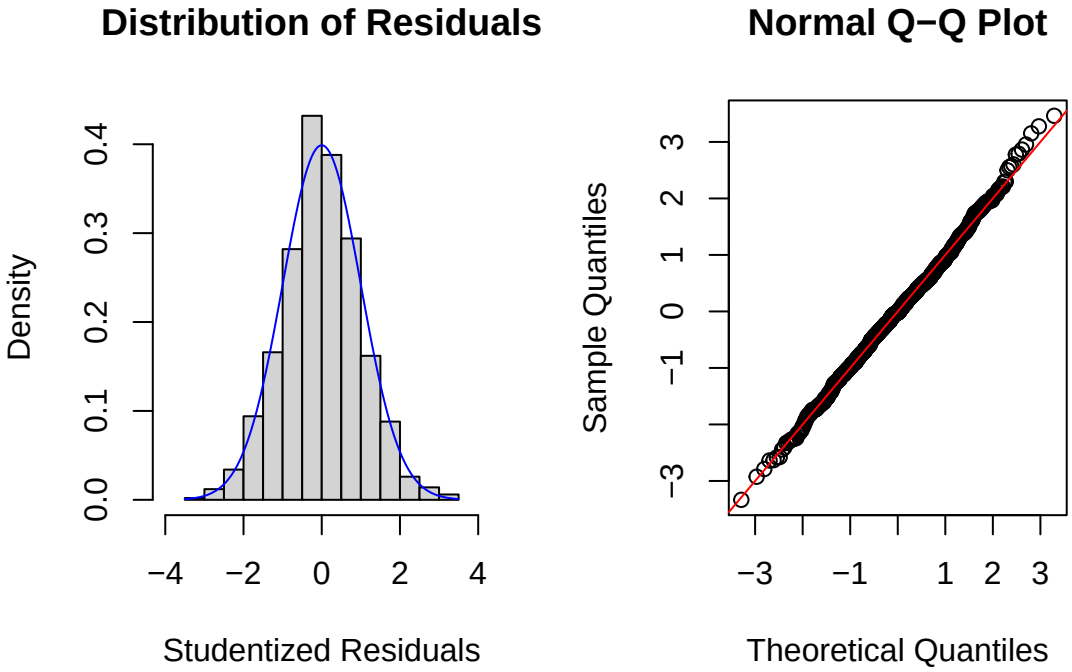


From the histogram we noticed that our outcome values has a fairly normal distribution.

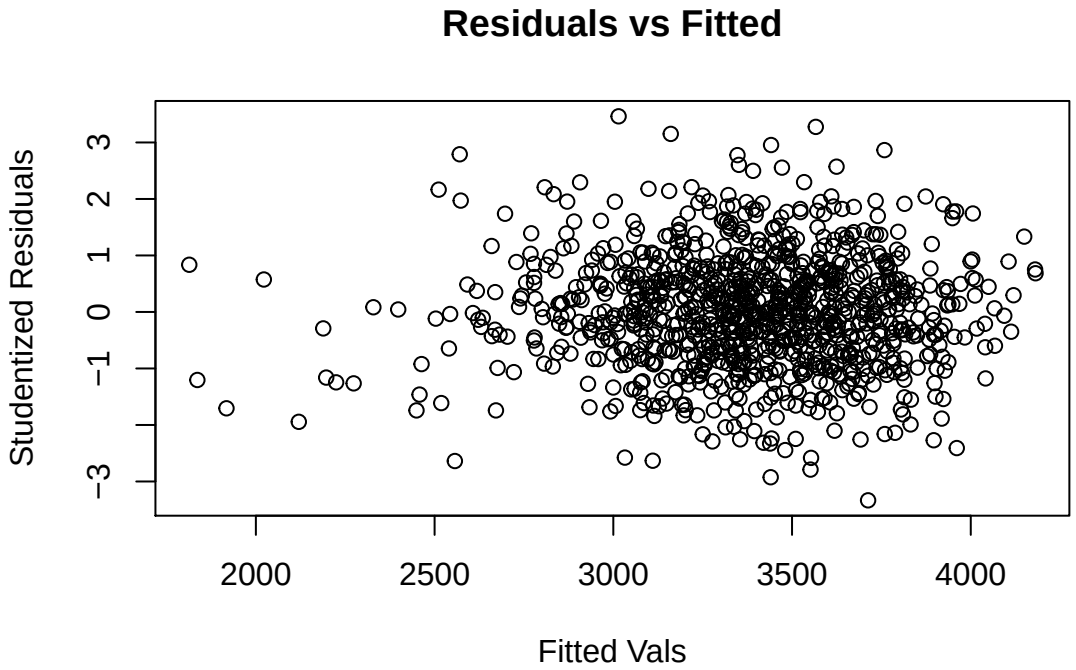
Methods

We first evaluated the different assumptions of multiple linear regression (i.e. normality, independence, linearity and homoskedasticity) for the full model where the birth weight was regressed over all 80 given covariates.





For linearity assumption, we used the *avplots()* function in R to assess the linearity of each covariates (above graphs only show the initial four plots as example, more details in appendix).To assess normality, we plotted a histogram of studentized residuals and a normal-QQ plot, the histogram had a distribution fairly similar to that of a $N(0, 1)$ distribution and most of the points on the QQ-plot lied on the 45 degree line, which indicated that the normal assumption was satisfied.



For homoskedasticity assumption, we plotted the studentized residuals against the fitted value. No mean-variance relationship was observed in the plot as the variability was fairly

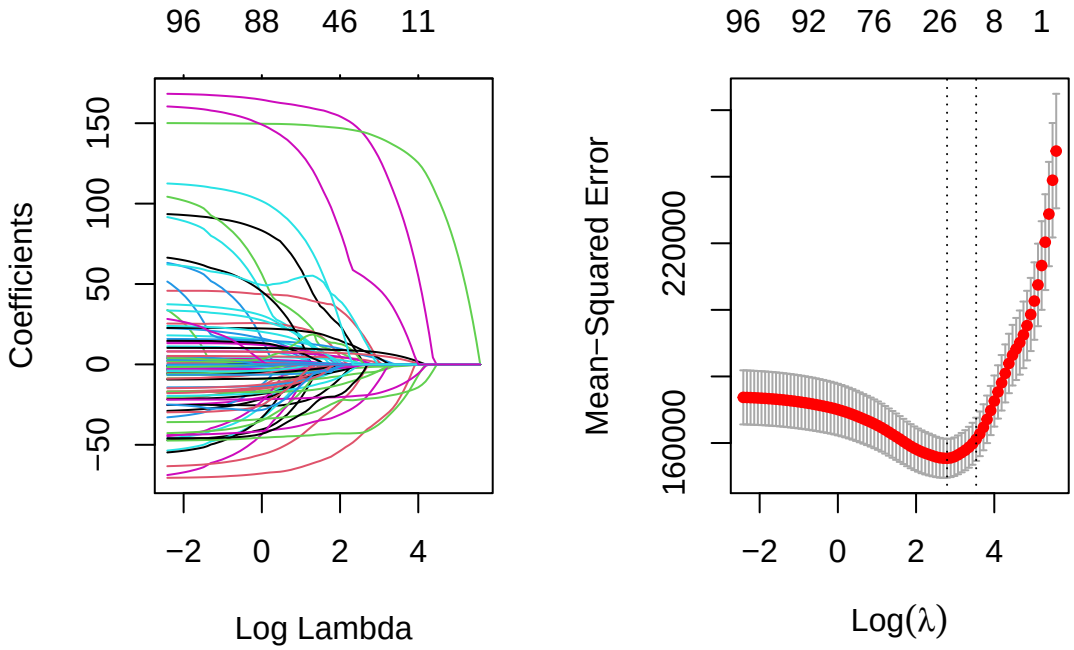
similar across all fitted values. Note that there is slightly less observed spread for lower fitted values (below 2500 g), but that can be explained by the lack of data. Overall, there was no observable pattern in the plot that would suggest heteroskedasticity. Then, among all the added variable plots, no significant pattern that would suggest non linearity was observed.The holded assumption of linearity suggested that a linear model might be used to model the birthweight to the other 80 covariates. Lastly, given that we were not given any information on data collection, we decided to assume that the independence assumption holds for the data.

For our statistical analysis, we first regressed the outcome of interest, birth weight, on all 80 covariates available in our full model. No transformation of the data was used as the basic assumptions of multiple linear regression were satisfied previously. Prior to model building, we investigated multicollinearity between the explanatory variables with the intention of removing all covariates with high multicollinearity. To do so, we used the *generalized variance inflation factor (GVIF)* of each covariates as a measure of multicollinearity, given that our full model contains both continuous and categorical variables. The reasoning behind this method is that categorical variables usually require more than 1 coefficient/degree of freedom, thus they are usually evaluated using GVIF. In order to compare the GVIF across dimensions (i.e compare between continuous and categorical data), we looked at the following transformation $GVIF^2(\frac{1}{2DF})$ which reduces the GVIF to a linear measure. Then, we squared the $GVIF^2(\frac{1}{2DF})$ value and applied the VIF rule of thumb (smaller than 10). As a result, we removed the covariate *h_humidity_preg_None (Humidity average during pregnancy)* from our full model as it was the only covariates with the square of $GVIF^2(\frac{1}{2DF})$ being greater than 10 (actual value 11.350). Now, the new model regressed over the remaining 79 covariates has Residual standard error: 396.8, Multiple R-squared: 0.4512, Adjusted R-squared: 0.3929.

Then, we used the new full model (with 79 covariates) and selected 3 different models, two using stepwise selection based on AIC and BIC and one using LASSO. In our first model, we used the new full model as the upper limit and the base model lm(e3_bw~1) as the lower limit for the stepwise selection, and used AIC as the criteria. This yielded a model with 20 covariates.

Similarly to our first model, our second model had similar set up with the only difference being that the criteria used was BIC. This in return gave us a smaller model containing 8 covariates.

Lastly, for our LASSO model, we used a training dataset of size 800 and testing dataset of size 200, both of which are randomly selected. Then, we applied the LASSO procedure to generate a model with 15 covariates.



Name of the model	Procedure Method
Full model	Contains all 80 covariates
New Full model	Delete the variable h_humidity_preg_None (Humidity average during pregnancy) with VIF larger or equal to 10 in full model
First model	Used AIC as the criteria to do the stepwise selection
Second model	Used BIC as the criteria to do the stepwise selection
Third model	LASSO procedure

the first model:

e3_bw ~ e3_gac_None + h_bro_preg_Log + e3_sex_None + h_mbmi_None + h_edumc_None + hs_wgtgain_None + hs_pfoa_m_Log2 + e3_asmokcigd_p_None + hs_dmtp_madj_Log2 + h_pm10_ratio_preg_None + hs_dep_madj_Log2 + hs_cs_m_Log2 + hs_mepa_madj_Log2 + hs_etpa_madj_Log2 + hs_pbde153_madj_Log2 + h_dairy_preg_Ter + h_meat_preg_Ter

p = 20, df = 979, Residual standard error: 391,
Multiple R-squared: 0.4222, Adjusted R-squared: 0.4104

second model:

e3_bw ~ e3_gac_None + h_bro_preg_Log + e3_sex_None + h_mbmi_None + hs_wgtgain_None + e3_asmokcigd_p_None + hs_pfoa_m_Log2 + hs_dmtp_madj_Log2

p = 8, df = 991, Residual standard error: 397.2,
Multiple R-squared: 0.3965, Adjusted R-squared: 0.3917

third model:

e3_bw ~ h_pm10_ratio_preg_None + h_dairy_preg_Ter + h_folic_t1_None +
hs_as_m_Log2 + hs_as_m_Log2 + hs_hg_m_Log2 + hs_pb_m_Log2+
hs_dmtp_madj_Log2 +
hs_pfoa_m_Log2 + hs_etpa_madj_Log2 + hs_mepa_madj_Log2 +
hs_mep_madj_Log2 +
hs_mibp_madj_Log2 + e3_asmokcigd_p_None + h_bro_preg_Log +
e3_sex_None + e3_yearbir_None + h_mbmi_None +
hs_wgtgain_None + e3_gac_None + h_edumc_None

p = 27, df = 972, Residual standard error:393.4,
Multiple R-squared: 0.4192, Adjusted R-squared: 0.403

The stepwise model built based on AIC had 20 parameters with adjusted r-squared 0.4104, the stepwise model built based on BIC had 8 parameters with adjusted r_squared 0.3917, and the model build based on LASSO and cross validation had 27 parameters with adjusetd r-sqaured 0.403. Since the adjusted r squared of the BIC model was close to the full model with 79 covariates, and it had the best parsimony(i.e smallest amount of covariates) among the 3 constructed models, we decided to use the second model, which is e3_bw ~ e3_gac_None + h_bro_preg_Log + e3_sex_None + h_mbmi_None + hs_wgtgain_None + e3_asmokcigd_p_None + hs_pfoa_m_Log2 + hs_dmtp_madj_Log2**, was the most appropriate one.

Furthermore, for our alternative goal, we built two additional models to check how chemical and non-chemical exposures influence the birthweight differently using stepwise selection with BIC as criteria. One model was about the influence of chemical exposure while the other one was about the influence of non-chemical exposure.

For the chemical exposure model, we first regressed the outcome of interest, birth weight, over the 50 variables related to chemical exposures, and set this as the upper limit for our stepwise selection. We then used the same base model lm(e3_bw~1) as previously as the lower limit. Using BIC as the criteria in stepwise selection, we obtained a model with 6 chemical exposure related covariates.

Similarly for non chemical exposure model, we regressed the outcome of interest over the 29 variables related to non-chemical exposures and set this as the upper limit for stepwise selection. Note that we did not include the covariates *h_humidity_preg_None* in the upper limit model as it was removed for high multicollinearity. Then, using the same base model as lower limit and BIC as criteria, we obtained a model with 6 non chemical exposure related variables.

Results

After model selection, we wanted to check if there were any outliers in the 1000 observations that may have effect on our BIC_model. we first found four different sets of high influential points using hatvalues, DFFITS, cook's distance and DFBETAS four methods, and tried to find a set of points that exist in all four sets. Then we omitted these observations to see if they have a large impact on BIC_model.

We first found points i with $h_i > 2\bar{h}$, which gave us the first set with 67 points. Next, we used DFFITS to measures the observation that has large impact on its fitted value, then we got the second set with 59 points that $|DFFITS_i| > 2\sqrt{\frac{p+1}{n}}$. We also used Cook's distance method to measure the observations' impact on all fitted values, but we got a third set containing zero element.Lastly, we use DFBETAS to measure i^{th} observations' impact on coefficient estimates, which gave us fourth set with 210 points. (more details of the points in each sets are in appendix).

There were 24 common points that exist in first, second and fourth set. They were (25 37 107 251 252 285 290 348 369 409 416 421 492 513 558 665 702 770 801 802 806 900 958 984).

Then, we wanted to check if there was a need to remove these high influential points, so we omitted these observations and made a new model to see if there was any difference.

The estimated coefficients and their corresponding p-value of the model without the 24 observations and BIC_model are shown in the following table (see appendix for estimated coeeficeints interpretation):

model omitted the 24 observation	estimates	p-value	residual standard error	adjusted r-squared
(Intercept)	-3463.597	< 2e-16	396.9	0.3924
e3_gac_None	162.687	< 2e-16		
h_bro_preg_Log	-34.833	3.12e-09		
e3_sex_Nonemale	161.463	4.23e-10		
h_mbmi_None	11.320	0.000165		
hs_wgtgain_None	7.741	0.000239		
hs_pfoa_m_Log2	-37.193	0.005097		
hs_dmtp_madj_Log2	11.125	0.005577		

BIC_model	estimates	p-value	residual standard error	adjusted r-squared
(Intercept)	-3471.787	< 2e-16	397.2	0.3917
e3_gac_None	163.102	< 2e-16		
h_bro_preg_Log	-34.875	1.83e-09		
e3_sex_Nonemale	157.725	6.54e-10		
h_mbmi_None	11.277	0.000159		
hs_wgtgain_None	7.728	0.000401		
hs_pfoa_m_Log2	-41.608	0.001181		

BIC_model	estimates	p-value	residual standard error	adjusted r-squared
hs_dmtp_madj_Log2	10.828	0.006493		

Since we had no information on data collection and the observed differences were no significant enough, we decided not to remove any observations.

For BIC_model, We could observe that all coefficients estimates in the model were significantly large, and each p-value for these eight covariates are far less than 0.05. So we rejected the null hypothesis that one of the coefficient will be zero. Thus, these eight covariates all have significant impact on birth weight. *e3_gac_None(Gestational age at birth (week))* is the variable that has the largest positive impact on *e3_bw (birth weight)*, *hs_pfoa_m_Log2(Perfluorooctanoate (PFOA) in mother)* is the variable that has the largest negative impact on *e3_bw (birth weight)*.

As for the chemical related exposure model, we found that there are three variables with a high coefficients, which meant that these three variables have the highest influence on the birthweight.

Chemical variable	Description	Coefficient
hs_pcb118_madj_Log2	Polychlorinated biphenyl-118 (PCB-118) in mother adjusted for lipids	91.655
hs_pcb153_madj_Log2	Polychlorinated biphenyl-153 (PCB-153) in mother adjusted for lipids	-55.863
hs_pfoa_m_Log2	Perfluorooctanoate (PFOA) in mother	-49.602

As for the non-chemical related exposure model, we found that there were two variables with a high coefficient, which meant these three variables have the highest influence on the birthweight.

Nonchemical variable	Description	Coefficient
e3_gac_None	Gestational age at birth (week)	160.243
e3_sex_None	Child sex (female / male)	147.376

Discussion

During the selection procedure, we built five models as the candidate models which has used AIC, BIC and LASSO. But we believe that our conclusions may not be totally accurate as the data provided may be biased. For example, people in different regions of the world

have different genetics and living habits, which makes people's physique very different, which may affect the birth weight. But the data provided did not include information about the women themselves. This makes it possible that our conclusions will be inaccurate. In addition, congenital conditions such as stunted growth and disabilities can affect the birthweight, but there is no significant evidence of these problems in the data. Lastly, given that no information were given to us with regards to how the samples were collected, we do not know if the independence assumption were really satisfied in our data, which could lead to errors in our models and conclusions.

Besides, we also found that when we built the models on chemical and non-chemical exposures' influence the birthweight separately, there were two variables that did not appear in the overall BIC model, which means that in the smaller models, these variables are important, but when they are in the overall model, they become unimportant.

Appendix

```
codebook <- read.csv("/Users/kaywu/Desktop/stat331/final\ project/codebook.csv")
load("/Users/kaywu/Desktop/stat331/final\ project/pollution.Rdata")
library(car)
# Scatterplot for chemical that has the highest coefficient (greatest
# influence):
## hs_pcb118_madj_Log2: 91.655
## hs_pcb153_madj_Log2: -55.863
## hs_pfoa_m_Log2: -49.602
plot(pollution$hs_pcb118_madj_Log2, pollution$e3_bw,
     main="Scatterplot Example",
     xlab="hs_pcb118_madj_Log2", ylab="e3_bw", pch=19)
# regression line (y~x)
abline(lm(pollution$e3_bw~pollution$hs_pcb118_madj_Log2), col="red")

# It follows almost log relationship. We can clearly see one outlier.
pcb118_outliers <- pollution[pollution$hs_pcb118_madj_Log2 > 6, ]
pcb118_outliers

# five number summary for hs_pcb118_madj_Log2:
summary(pollution$hs_pcb118_madj_Log2)

# hs_pcb153_madj_Log2 has a 2nd high coefficient but forward selection does
#  not select it into the whole model, only the chemical model.
plot(pollution$hs_pcb153_madj_Log2, pollution$e3_bw,
     main="Scatterplot Example", xlab="hs_pcb153_madj_Log2",
     ylab="e3_bw", pch=19)
# regression line (y~x)
abline(lm(pollution$e3_bw~pollution$hs_pcb153_madj_Log2),
       col="red")


#We notice that line 289 appear twice, but it is not in the outlier list.
#  Five number summary for hs_pcb153_madj_Log2:
summary(pollution$hs_pcb153_madj_Log2)

### hs_pfoa_m_Log2
plot(pollution$hs_pfoa_m_Log2, pollution$e3_bw, main="hs_pfoa_m_Log2",
     xlab="hs_pfoa_m_Log2", ylab="e3_bw", pch=19)
# regression line (y~x)
abline(lm(pollution$e3_bw~pollution$hs_pfoa_m_Log2), col="red")

# We can also see 2 outliers.
pfoa_outliers <- pollution[pollution$hs_pfoa_m_Log2 < -4, ]
```

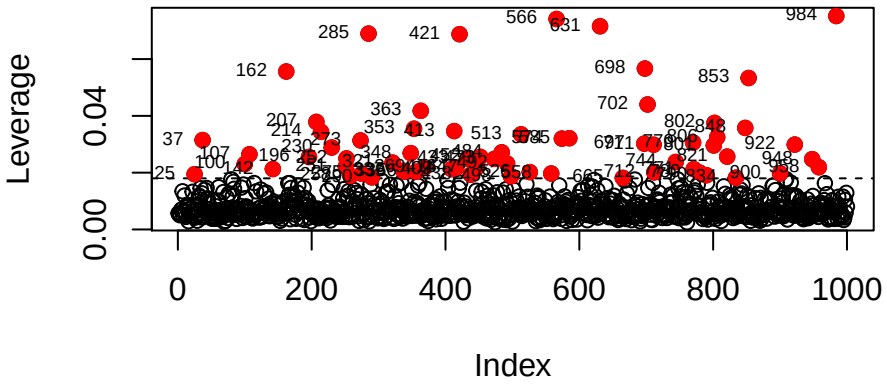
```r
pfoa_outliers

# 285 and 698 are both points that has high leverage. Five number summary
#  for hs_pfoa_m_Log2:
summary(pollution$hs_pfoa_m_Log2)

# We have also one outlier.
pcb153_outliers <- pollution[pollution$hs_pcb153_madj_Log2 > 8, ]
pcb153_outliers

## Non-chemical:
# e3_gac_None: 160.243
# e3_sex_Nonemal: 147.376

plot(pollution$e3_gac_None, pollution$e3_bw, main="e3_gac_None",
     xlab="e3_gac_None", ylab="e3_bw", pch=19)
# regression line (y~x)
abline(lm(pollution$e3_bw~pollution$e3_gac_None), col="red")

# e3_gac_None has a clearly linear + log relationship between birthweight.
plot(pollution_NonChem$e3_sex_None, pollution_NonChem$e3_bw,
     main="e3_sex_None",xlab="e3_sex_None", ylab="e3_bw", pch=19)

#custom ordering for column 16 & 27 (just for a more logical ordering of
#  categories & will make interpretation slightly easier)
exercisefreq <- c("Low", "Medium", "High")
hs_tl_mdich_levels <- c("Undetected", "Detected")
#replace in dataset with the new ordering
h_pavig_t3_None2<- factor(pollution$h_pavig_t3_None, levels = exercisefreq)
pollution$h_pavig_t3_None<-h_pavig_t3_None2
hs_tl_mdich_None2<-factor(pollution$hs_tl_mdich_None,
                          levels = hs_tl_mdich_levels)
pollution$hs_tl_mdich_None<-hs_tl_mdich_None2
Mfull<-lm(e3_bw~.,data=pollution)
## residuals
# raw residuals
res1 <- resid(Mfull)
# studentized residuals
stud1 <- res1/(sigma(Mfull)*sqrt(1-hatvalues(Mfull)))

## partial regression (aka added variable plots)
library(OpenImageR)
im <- readImage("/Users/kaywu/Desktop/aovplot.png")
imageShow(im)
```

```r
par(mfrow = c(1,2))
# normality
hist(stud1,breaks=12,
     probability=TRUE,xlim=c(-4,4),
     xlab="Studentized Residuals",
     main="Distribution of Residuals")
grid <- seq(-3.5,3.5,by=0.05)
lines(x=grid,y=dnorm(grid),col="blue")
qqnorm(stud1)
abline(0,1,col="red")

#homoskedasticity
## plot of studentized residuals vs fitted values
par(mfrow = c(1,1))
plot(stud1~fitted(Mfull),
     xlab="Fitted Vals",
     ylab="Studentized Residuals",
     main="Residuals vs Fitted")

############################################################
# multicollinearity
# reference: http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_
#   Book/4-5-Multiple-collinearity.html
# essentially, the built in vif() function will output GVIF^(1/(2*Df)),
#    we need to square this value and apply the usual rule of thumb for
#  VIF<10
#    recursive function that takes in a response variable y and a dataset,
#  and removes a covariate with the highest GVIF/VIF > 10,
#  Then recurse on itself until all GVIF/VIF
#    in the remaining model is smaller than 10, y must be in string form
#  dataset must only contain the response variable in the first column and
#  the desired covariates after
remove_high_GVIF <- function(y, dataset){
    # list all variables in dataset
    coefficients <- variable.names(dataset)
    # removes the response variable from coefficients
    coefficients <- coefficients[-1]

    #base model with y regressed on covariates in coefficients
    full_model<-lm(as.formula(paste(y,"~",paste(coefficients, collapse=" + "))),
                   data=dataset)
    #  returns GVIF^(1/(2*Df)) squared to which we can apply the
    #  rule of thumb for VIF>10
    gvifs<-vif(full_model)[,3]^2
```

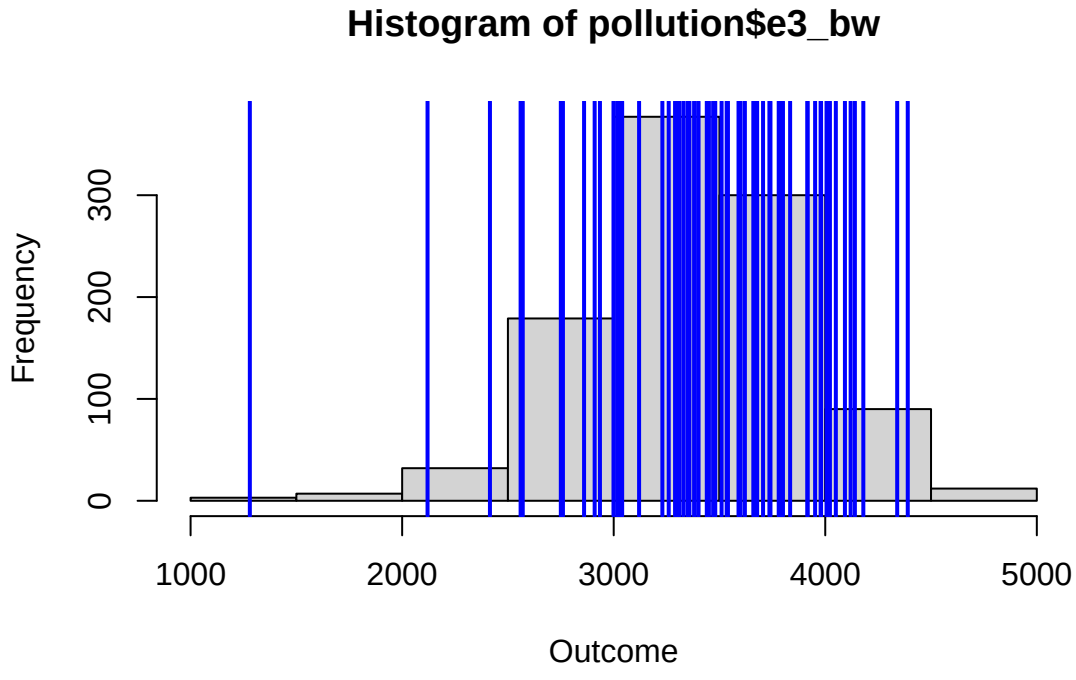
```
    # returns the highest GVIF/VIF in our model
    highest<-max(gvifs)
    #if our highest GVIF/VIF is lower than 10, return that model
    if(highest < 10){return(full_model)}
    else{   # returns the index for the the highest GVIF/VIF
            ind <- match(max(gvifs),gvifs)
            # removes the column corresponding to the coefficient
            #  with the highest VIF in our dataset
            new_dataset <- dataset[-(ind+1)]
            # recurse on the function with the new dataset
            remove_high_GVIF(y,new_dataset)
        }
}

#the model after removing high multicollinearity
new_model<-remove_high_GVIF("e3_bw",pollution)
# only h_humidity_preg_None was removed from mfull for high
#  multicollinearity with GVIF of 11.349954
# new_model has p = 96, df = 903, Residual standard error: 396.8,
# Multiple R-squared:  0.4512, Adjusted R-squared:  0.3929
# model selection (stepwise)
# use AIC and BIC to make model selection
n <- nrow(pollution)
M0 <-lm(e3_bw~1, data=pollution)
AIC_model<-step(object = M0,
                scope = list(lower = M0, upper = new_model),
                direction = "both", trace = 1, k=2)
summary(AIC_model)
# the first model:
#e3_bw ~ e3_gac_None + h_bro_preg_Log + e3_sex_None + h_mbmi_None +
#     h_edumc_None + hs_wgtgain_None + hs_pfoa_m_Log2 + e3_asmokcigd_p_None
#     + hs_dmtp_madj_Log2 + h_pm10_ratio_preg_None + hs_dep_madj_Log2 +
#     hs_cs_m_Log2 + hs_mepa_madj_Log2 + hs_etpa_madj_Log2
#     + hs_pbde153_madj_Log2 + h_dairy_preg_Ter + h_meat_preg_Ter
#  AIC_model has p = 20, df = 979, Residual standard error: 391,
#   Multiple R-squared:  0.4222, Adjusted R-squared:  0.4104
BIC_model<-step(object = M0,
                scope = list(lower = M0, upper = new_model),
                direction = "both", trace = 1, k=log(n))
summary(BIC_model)
# second model:
# e3_bw ~ e3_gac_None + h_bro_preg_Log + e3_sex_None + h_mbmi_None +
#  hs_wgtgain_None + e3_asmokcigd_p_None + hs_pfoa_m_Log2
#  + hs_dmtp_madj_Log2
```

```r
# BIC_model has p = 8, df = 991, Residual standard error: 397.2,
#   Multiple R-squared:  0.3965, Adjusted R-squared:  0.3917

#LASSO
library(glmnet)
X <- model.matrix(new_model)[,-1] ## get covariates
y <- pollution$e3_bw  ## get outcome
pollution <- pollution[sample(nrow(pollution)),]
#800 in train dataset and 200 in test dataset
ntrain <- 800
train_id <- 1:ntrain
X_train <- X[train_id,]
X_test <- X[-train_id,]
y_train <- y[train_id]
y_test <- y[-train_id]

## fit models
M_lasso <- glmnet(x=X_train,y=y_train,alpha = 1)

## plot paths
plot(M_lasso,xvar = "lambda",label=TRUE)

## fit with crossval
cvfit_lasso <-  cv.glmnet(x=X_train,y=y_train,alpha = 1)

## plot MSPEs by lambda
plot(cvfit_lasso)

## estimated betas for minimum lambda
beta_lasso <- coef(cvfit_lasso, s = "lambda.min")
# the number of parameters in the model fitted using LASSO
par_length_lasso  <- length(beta_lasso@i)
r_squared_lasso <- 1 - min(cvfit_lasso$cvm) / var(pollution$e3_bw)
# par_length_lasso = 21, r_squared_lasso = 0.4008314

############################################################
# interpretation of beta in the BIC_model
#     1. the estimated Child weight at birth is -3471.787 g when all other
# covariates are zero.
#     2. holding other covariates unchanged, the estimated Child weight at
# birth is estimated to increase 163.102 g for one week increase in
# Gestational age at birth;
#     3. holding other covariates unchanged, the estimated Child weight at
# is estimated to decrease 34.875 g for one unit increase in Total
```

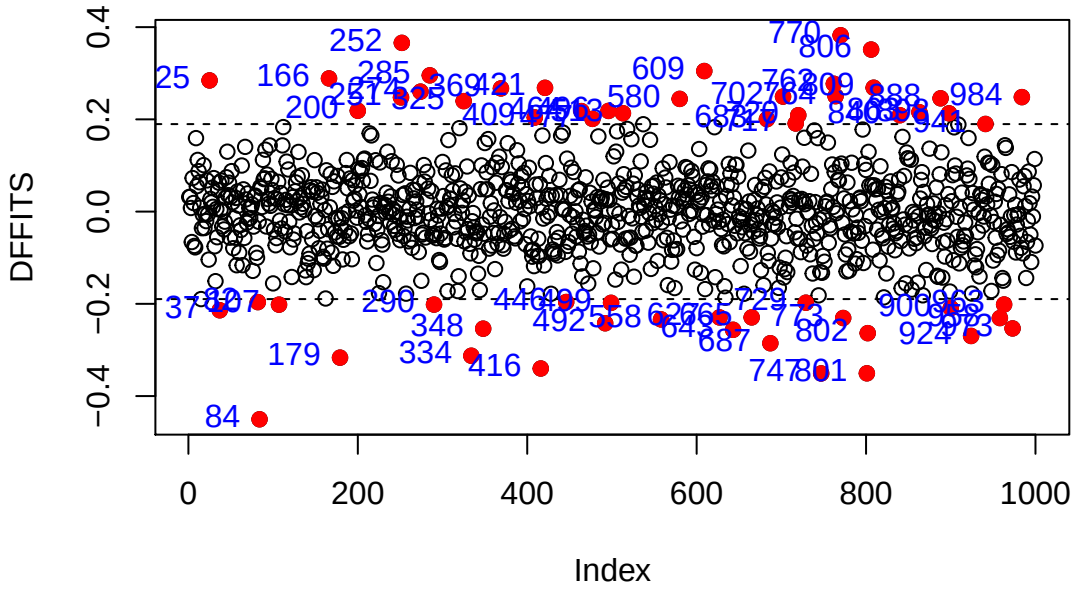
```r
# concentration of Brominated
# during pregnancy;
#     4. holding other covariates unchanged, the estimated Child weight at
# birth is estimated to increase 157.725 g for male child than
# female child;
#     5. holding other covariates unchanged, the estimated Child weight at
# birth is estimated to increase 11.277 g for one kg/m2 increase of
# Maternal pre-pregnancy body mass index;
#     6. holding other covariates unchanged, the estimated Child weight
# at birth is estimated to increase 7.728 g for one kg increase of maternal
# weight gain during pregnancy;
#     7. holding other covariates unchanged, the estimated Child weight at
# birth is estimated to decrease 24.755 g for one unit increase of maternal
# active Tobacco Smoke pregnancy mean nb cig/day;
#     8. holding other covariates unchanged, the estimated Child weight at
# birth is estimated to decrease 41.608 g for one unit increase of
# Perfluorooctanoate (PFOA) in mother;
#     9. holding other covariates unchanged, the estimated Child weight
# at birth is estimated to increase 10.828 g for one unit increase of
# Dimethyl thiophosphate (DMTP) in child adjusted for creatinine.


# detecting outliers
# Leverage
M <- BIC_model
 ## leverage (h_i)
lev <- hatvalues(M)
## \bar{h}
hbar <- mean(lev)
## plot leverage
plot(lev,ylab="Leverage")
## add line at 2hbar
abline(h=2*hbar,lty=2)
## x values for labelling points >2hbar
first_set <- which(lev>2*hbar)
 ## add red points >2hbar
points(lev[first_set]~first_set,col="red",pch=19)
## label points >2hbar
text(x=first_set,y=lev[first_set], labels=first_set, cex= 0.6, pos=2)
```




```
# there are 67 observation with leverge larger than 2(\bar{h})
# first_set = [25  37 100 107 142 162 196 207 214 230 251 252 258 273 275
# 285 290 321 337 339 348 353 25  37 100 107 142 162 196 207 214 230
# 251 252 258 273 275 285 290 321 337 339 348 353 356 363 369 409 413
# 416 421 433 438 447 452 473 484 492 498 513 526 558 566 574 585 631
# 356 363 369 409 413 416 421 433 438 447 452 473 484 492 498 513 526
# 558 566 574 585 631 665 697 698 702 711 712 744 770 771 780 790 801
# 802 806 821 834 848 853 900 922 948 958 665 697 698 702 711 712 744
# 770 771 780 790 801 802 806 821 834 848 853 900 922 948 958 984]
# the highest leverage points is observation 984.

## is it a y-outlier?  (figure it later)
hist(pollution$e3_bw,xlab="Outcome")
## add highest leverage point
abline(v=pollution$e3_bw[first_set],lwd=2,col="blue")
```



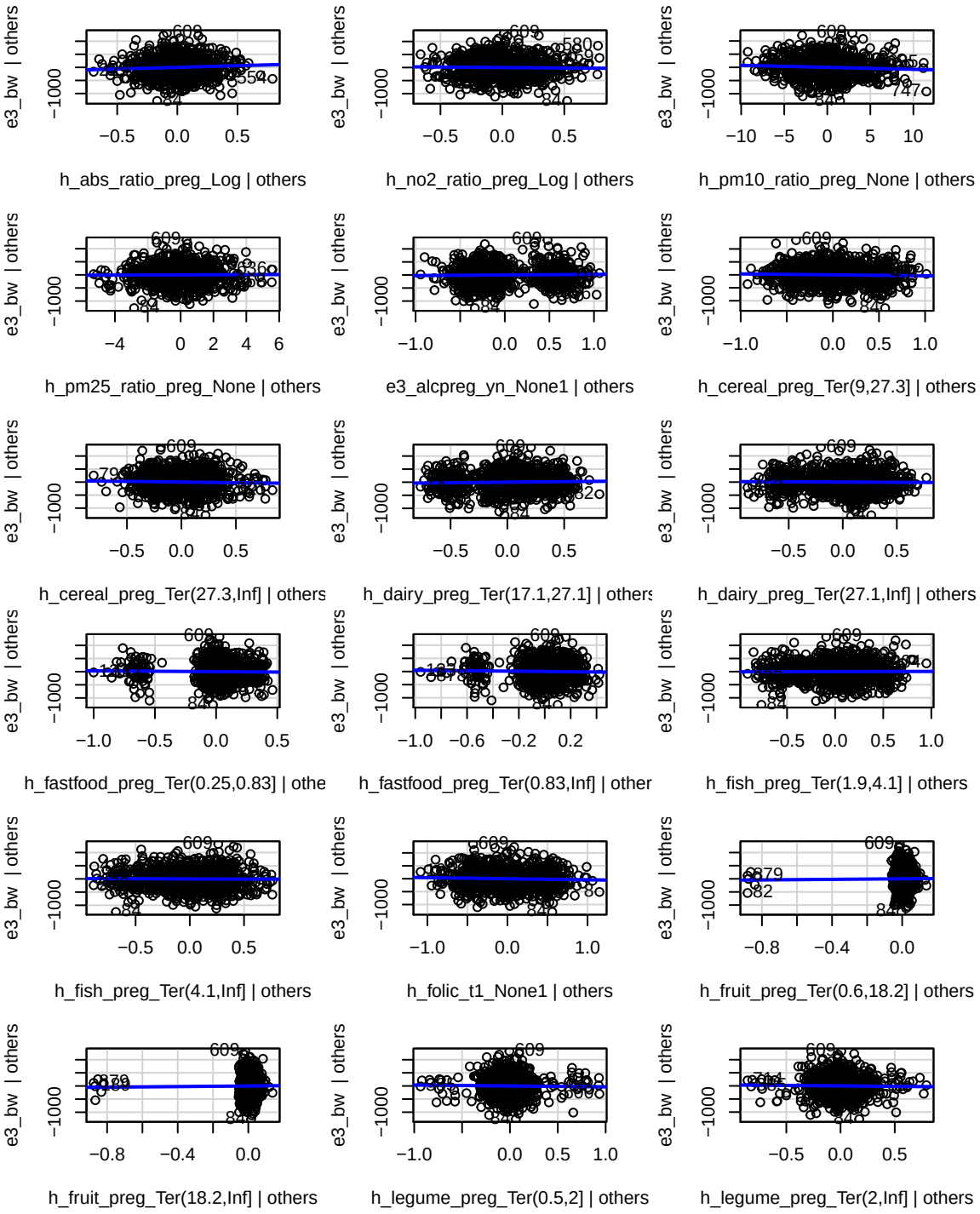

```r
## the graph shows the high leveraged observations spread almost all
# range in e3_bw so we think we do not need to remove the high leverage
# points.
# DFFITS (measures ith observation's impact on its fitted value)
# number of parameters in BIC_model
p <- summary(BIC_model)$fstatistic[2]
dffits_m <- dffits(M)
## plot DFFITS
plot(dffits_m,ylab="DFFITS")
abline(h=2*sqrt((p+1)/n),lty=2)  ## add thresholds
abline(h=-2*sqrt((p+1)/n),lty=2)
## highlight influential points
second_set <- which(abs(dffits_m)>2*sqrt((p+1)/n))
points(dffits_m[second_set]~dff_ind,col="red",pch=19) ## add red points
## label high influence points
text(y=dffits_m[second_set],x=dff_ind, labels=dff_ind, pos=2,col="blue")
```

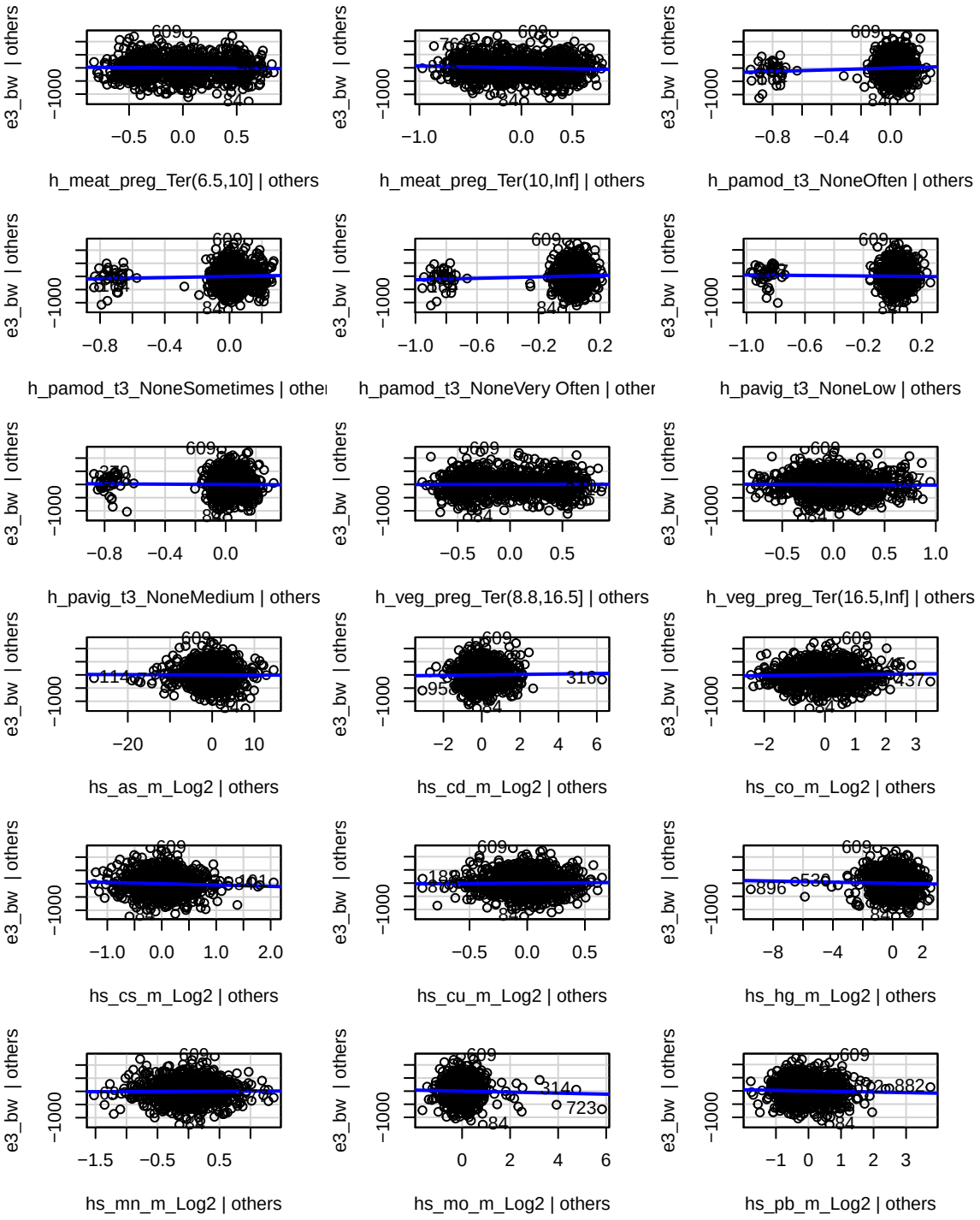


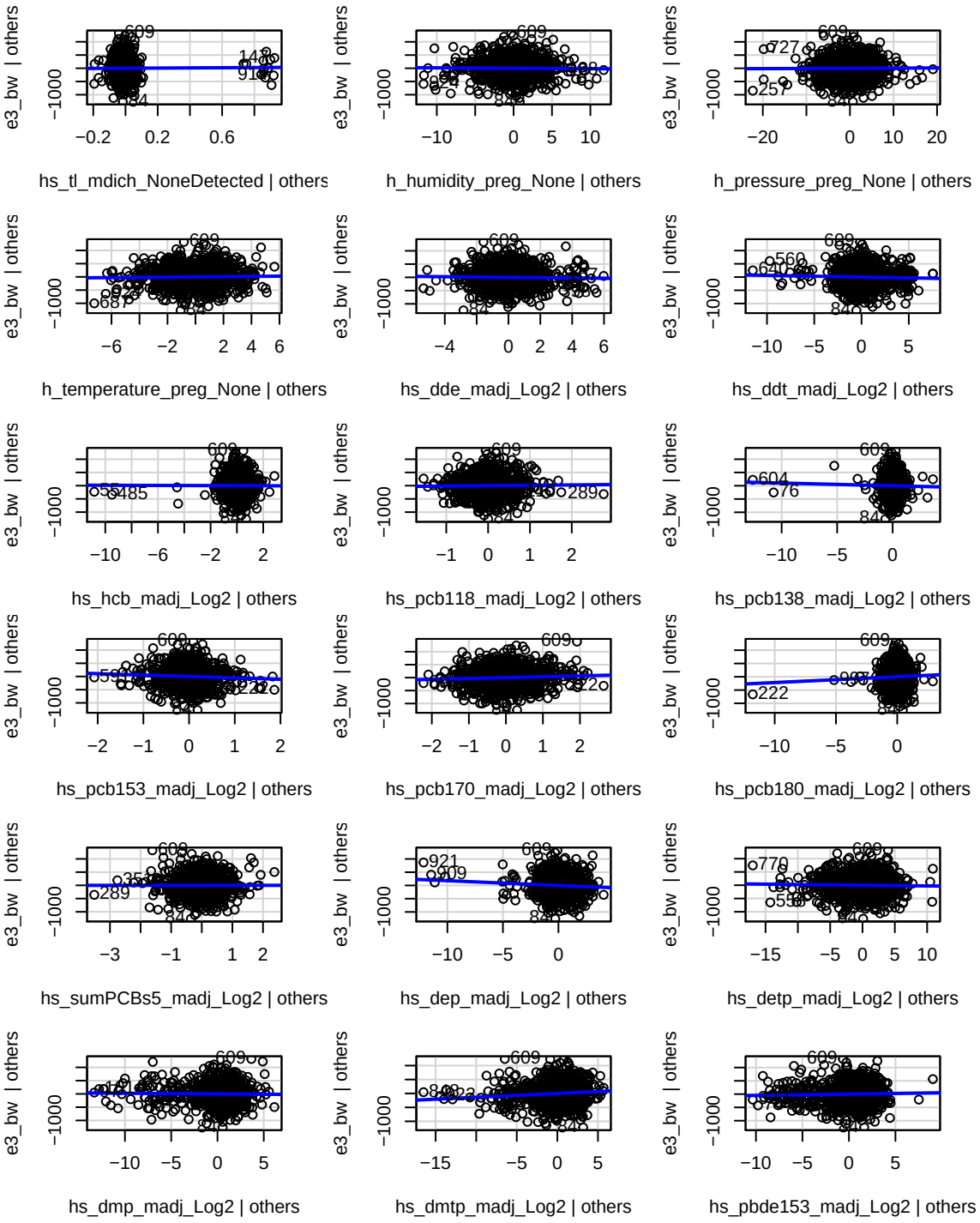

```r
# dff_ind has 59 influential points, they are
# 25  37  82  84 107 166 179 200 251 252 274 285 290 325 334 348 369
# 409 416 421 446 464 477 492 496 499 25  37  82  84 107 166 179 200
# 251 252 274 285 290 325 334 348 369 409 416 421 446 464 477 492 496
# 499 513 558 580 609 627 643 665 683 687 702 717 720 729 747 762 764
# 770 773 801 802 806 809 840 863 888 898 513 558 580 609 627 643 665
# 683 687 702 717 720 729 747 762 764 770 773 801 802 806 809 840 863
# 888 898 900 924 941 958 963 973 984 900 924 941 958 963 973 984

# cook's distance (measured ith observations's impact on all fitted values)
D <- cooks.distance(M) # Cook's distance
## influential points
```

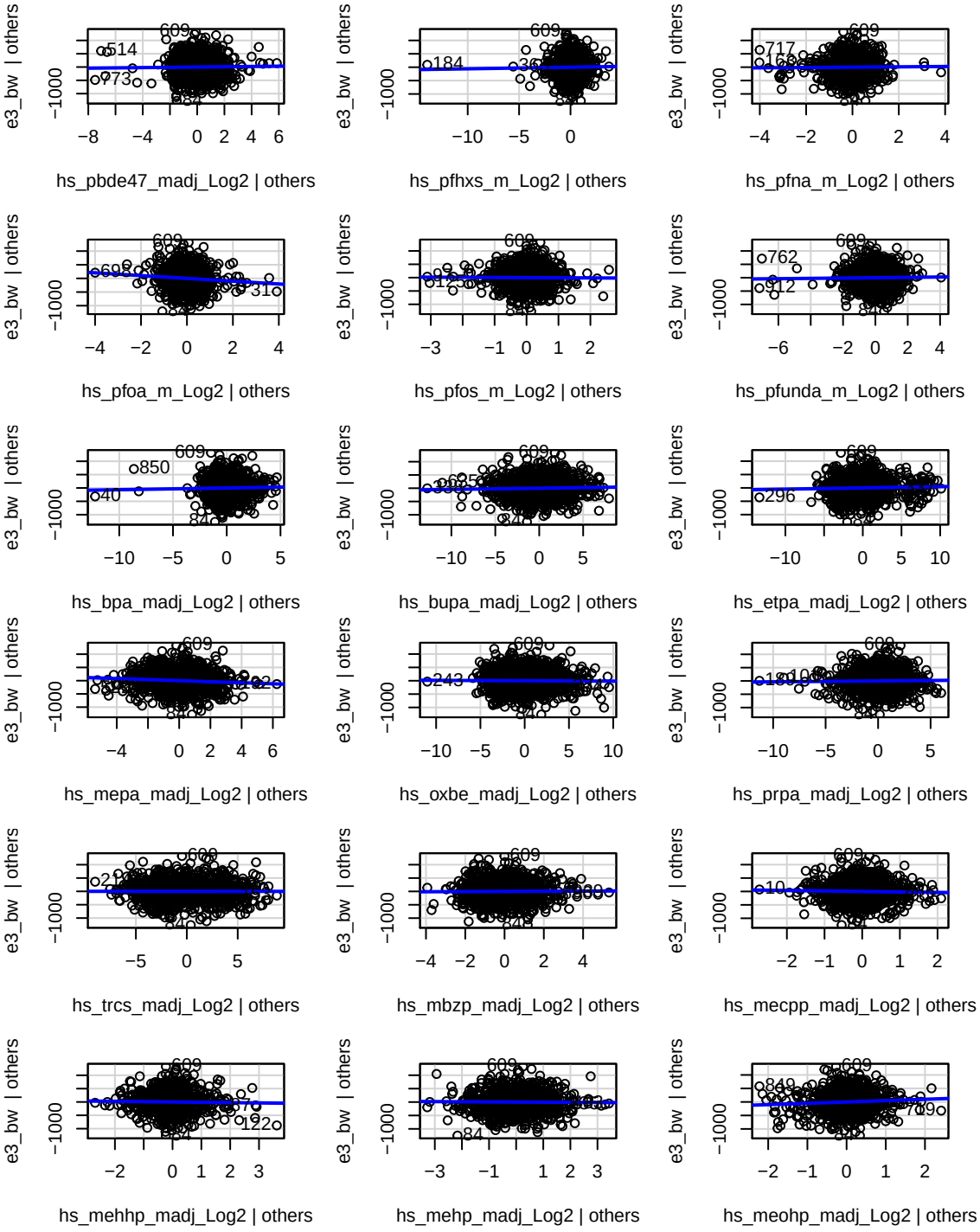
```r
third_set <- which(pf(D,p+1,n-p-1,lower.tail=TRUE)>0.05)
#   the inf_ind is empty, so under cook's distance method,
#   no observations in pollution that has influential impact
#   on all fitted values.

#DFBetas (measures ith observations'impact on coefficients estimates)
DFBETAS <- dfbetas(M)
ncol <- dim(DFBETAS)

fourth_set <- which(abs(DFBETAS[,2])>2/sqrt(n))
for (i in 3:ncol[2]){
  fourth_set <- unique(append(fourth_set, which(abs(DFBETAS[,i])>2/sqrt(n))))
}

# the intersection of first_set(levrage,jackknife) and second_set(DFFITS)
intersections <- intersect(first_set, intersect(fourth_set, second_set))
# [25  37 107 251 252 285 290 348 369 409 416 421 492 513 558 665 702
# 770 801 802 806 900 958 984]
# length is 24
# observation 984 has the highest leverage of all observations,


# for BIC_model, e3_gac_None(Gestational age at birth (week)), has the
# largest absolute coefficient value  163.102, so the most impacted
# variable is e3_gac_None.


############################################################
###### try to omit intersections
omit_ind <- intersections
M_omit <- lm(e3_bw ~ e3_gac_None + h_bro_preg_Log + e3_sex_None +
               h_mbmi_None + hs_wgtgain_None + e3_asmokcigd_p_None
             + hs_pfoa_m_Log2 + hs_dmtp_madj_Log2,
       data = pollution[-omit_ind,])
summary(M_omit)
pred_omit <- predict(M_omit,newdata = pollution)
############################################################

###### build model between chemical variables and non chemical variables.
# we separate into two dataframes for analyze
data_Chem = codebook[which(codebook$domain=='Chemicals'), ]
data_NonChem = codebook[which(codebook$domain!='Chemicals'), ]
ChemName = data_Chem$variable_name
NonChemName = data_NonChem$variable_name
```

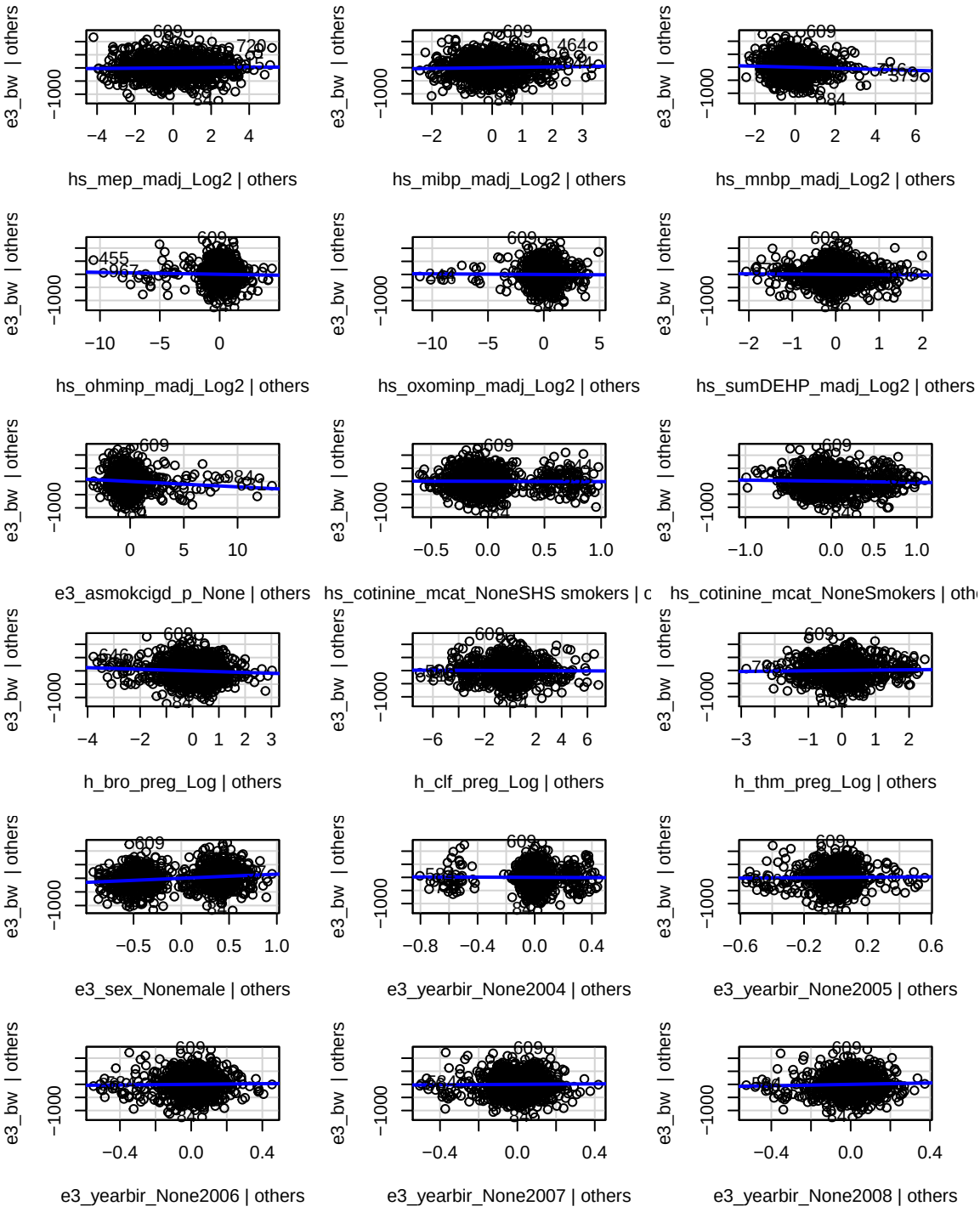
```r
library(dplyr)
pollution_Chem = select(pollution, ChemName)
pollution_Chem$e3_bw=pollution$e3_bw
pollution_NonChem = select(pollution, NonChemName)

# for chemical
Mfull_chem<-lm(e3_bw~., data=pollution_Chem) # build model with BIC
n_chem <- nrow(pollution_Chem)
M0_chem <-lm(e3_bw~1, data=pollution_Chem)

BIC_model_chem<-step(object = M0_chem,
                scope = list(lower = M0_chem, upper = Mfull_chem),
                direction = "both", trace = 1, k=log(n_chem))

# for non chemical

Mfull_Nonchem<-lm(e3_bw~., data=pollution_NonChem)
n_Nonchem <- nrow(pollution_NonChem)
M0_Nonchem <-lm(e3_bw~1, data=pollution_NonChem)

BIC_model_Nonchem<-step(object = M0_Nonchem,
                scope = list(lower = M0_Nonchem,
                             upper = Mfull_Nonchem),
                direction = "both", trace = 1, k=log(n_Nonchem))
```

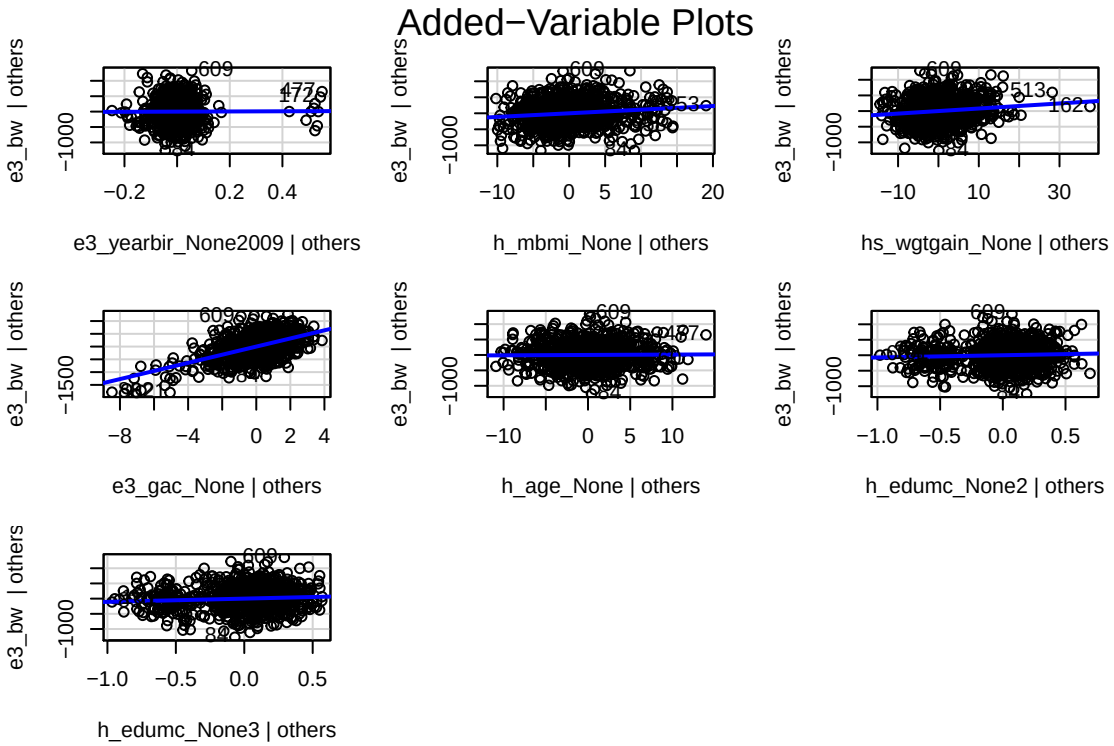

e3_bw | others
−1000
h_abs_ratio_preg_Log | others
h_no2_ratio_preg_Log | others
h_pm10_ratio_preg_None | others
h_pm25_ratio_preg_None | others
e3_alcpreg_yn_None1 | others
h_cereal_preg_Ter(9,27.3] | others
h_cereal_preg_Ter(27.3,Inf] | others
h_dairy_preg_Ter(17.1,27.1] | others
h_dairy_preg_Ter(27.1,Inf] | others
h_fastfood_preg_Ter(0.25,0.83] | othe
h_fastfood_preg_Ter(0.83,Inf] | other
h_fish_preg_Ter(1.9,4.1] | others
h_fish_preg_Ter(4.1,Inf] | others
h_folic_t1_None1 | others
h_fruit_preg_Ter(0.6,18.2] | others
h_fruit_preg_Ter(18.2,Inf] | others
h_legume_preg_Ter(0.5,2] | others
h_legume_preg_Ter(2,Inf] | others


e3_bw | others
h_meat_preg_Ter(6.5,10] | others
h_meat_preg_Ter(10,Inf] | others
h_pamod_t3_NoneOften | others
h_pamod_t3_NoneSometimes | other
h_pamod_t3_NoneVery Often | other
h_pavig_t3_NoneLow | others
h_pavig_t3_NoneMedium | others
h_veg_preg_Ter(8.8,16.5] | others
h_veg_preg_Ter(16.5,Inf] | others
hs_as_m_Log2 | others
hs_cd_m_Log2 | others
hs_co_m_Log2 | others
hs_cs_m_Log2 | others
hs_cu_m_Log2 | others
hs_hg_m_Log2 | others
hs_mn_m_Log2 | others
hs_mo_m_Log2 | others
hs_pb_m_Log2 | others


e3_bw | others
hs_tl_mdich_NoneDetected | others
h_humidity_preg_None | others
h_pressure_preg_None | others
h_temperature_preg_None | others
hs_dde_madj_Log2 | others
hs_ddt_madj_Log2 | others
hs_hcb_madj_Log2 | others
hs_pcb118_madj_Log2 | others
hs_pcb138_madj_Log2 | others
hs_pcb153_madj_Log2 | others
hs_pcb170_madj_Log2 | others
hs_pcb180_madj_Log2 | others
hs_sumPCBs5_madj_Log2 | others
hs_dep_madj_Log2 | others
hs_detp_madj_Log2 | others
hs_dmp_madj_Log2 | others
hs_dmtp_madj_Log2 | others
hs_pbde153_madj_Log2 | others


e3_bw | others
−1000
hs_pbde47_madj_Log2 | others
hs_pfhxs_m_Log2 | others
hs_pfna_m_Log2 | others
hs_pfoa_m_Log2 | others
hs_pfos_m_Log2 | others
hs_pfunda_m_Log2 | others
hs_bpa_madj_Log2 | others
hs_bupa_madj_Log2 | others
hs_etpa_madj_Log2 | others
hs_mepa_madj_Log2 | others
hs_oxbe_madj_Log2 | others
hs_prpa_madj_Log2 | others
hs_trcs_madj_Log2 | others
hs_mbzp_madj_Log2 | others
hs_mecpp_madj_Log2 | others
hs_mehhp_madj_Log2 | others
hs_mehp_madj_Log2 | others
hs_meohp_madj_Log2 | others


e3_bw | others
hs_mep_madj_Log2 | others
hs_mibp_madj_Log2 | others
hs_mnbp_madj_Log2 | others
hs_ohminp_madj_Log2 | others
hs_oxominp_madj_Log2 | others
hs_sumDEHP_madj_Log2 | others
e3_asmokcigd_p_None | others
hs_cotinine_mcat_NoneSHS smokers | c
hs_cotinine_mcat_NoneSmokers | oth
h_bro_preg_Log | others
h_clf_preg_Log | others
h_thm_preg_Log | others
e3_sex_Nonemale | others
e3_yearbir_None2004 | others
e3_yearbir_None2005 | others
e3_yearbir_None2006 | others
e3_yearbir_None2007 | others
e3_yearbir_None2008 | others


Added−Variable Plots
e3_bw | others
e3_yearbir_None2009 | others
h_mbmi_None | others
hs_wgtgain_None | others
e3_gac_None | others
h_age_None | others
h_edumc_None2 | others
h_edumc_None3 | others