LR Project1

Kuan Teh Wan

April 2024

1 Part 1(a)

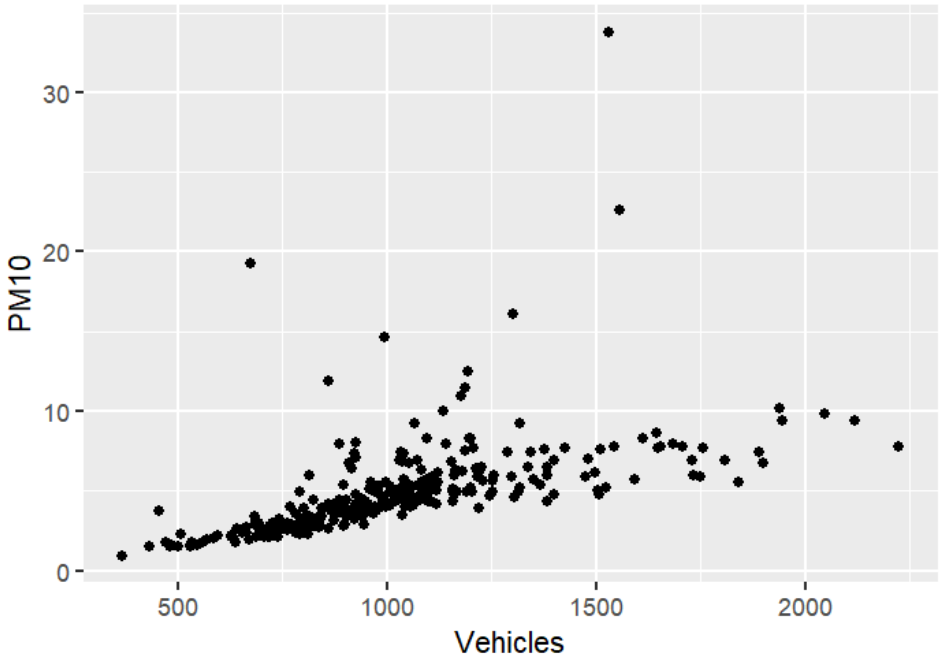


Figure 1: PM10/Vehicles plot

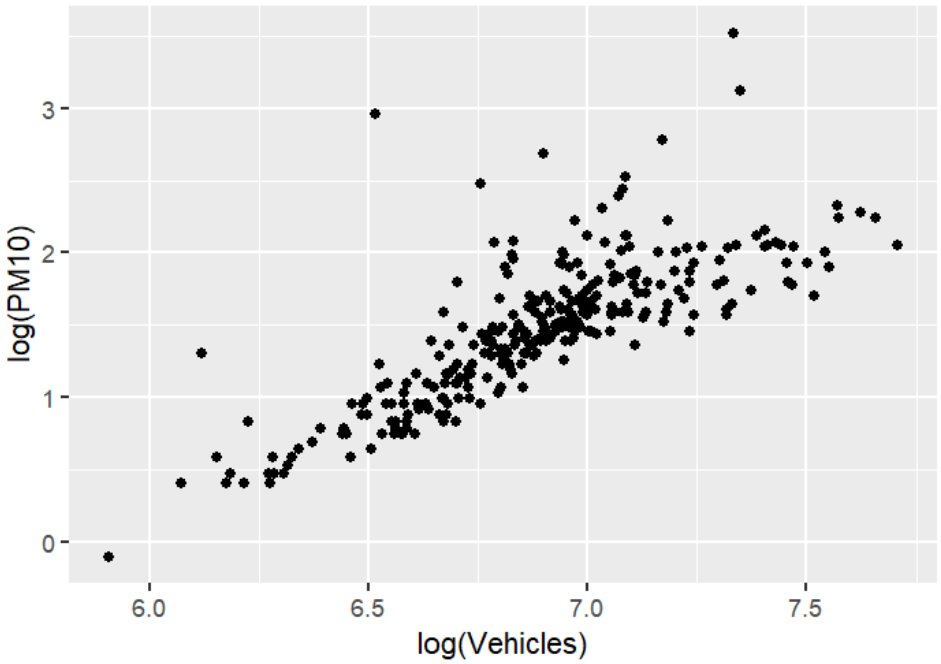


Figure 2: log(PM10)/log(Vehicles plot)

From the figures above we can see that, we should take the logarithm of PM10 since it significantly reduces the variability of the model and also it helps to transform the skewed PM10 data(most are small, only a few are large) into a more symmetric manner. Another point worth mentioning is that, Logarithmic transformations can help linearize relationships between variables, making it easier to apply linear techniques later on.

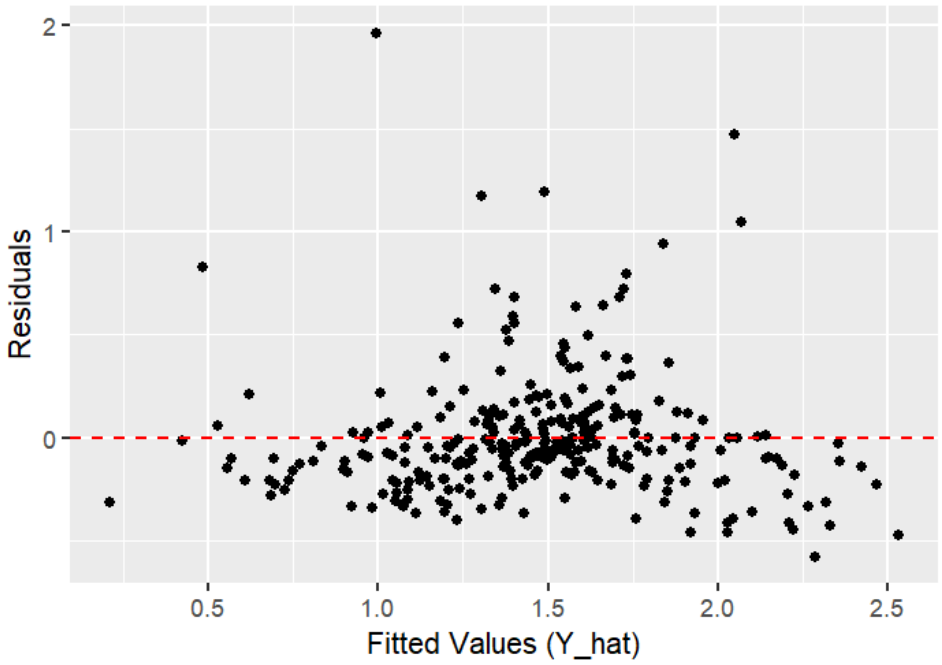


Figure 3: log(PM10)/log(Vehicles) residual plot

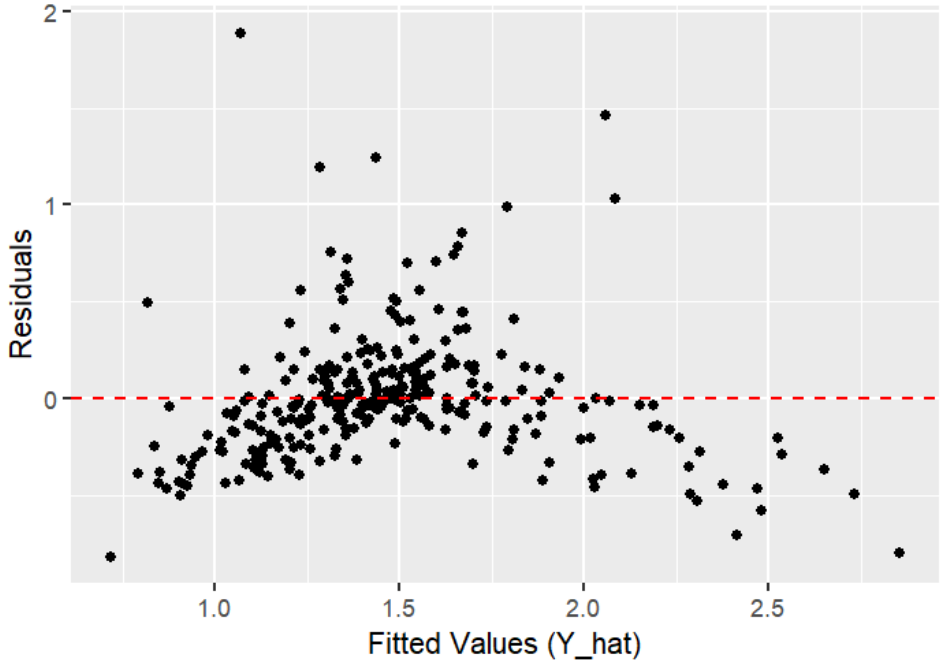


Figure 4: log(PM10)/Vehicles residual plot

From another perspective, let's take a look at the residual plots to decide whether to use x=Vehicles or x=log(Vehicles). In figure 3, we can see that data points and are clustered more around a red dashed line at 0 residual when compared to figure 4 and this indicates a better fit for the model.

1.(b)

In this subsection, we discussed the modeling $ln(PM10) = \beta_0 + \beta_1 x + \epsilon$, and two plots which depicted the relationship of $ln(PM10)$ and $ln(Vehicles)$ respectively. To start with, the $\beta-$estimates with 95% confidence intervals are shown below:

Param List	Estimate	95% COnfidence Interval
β_0	-7.389	[-8.194491,-6.583741]
β_1	1.287	[1.170085,1.403780]

Table 1: The $\beta-$estimates and 95% confidence intervals

After that, the plot with $ln(PM10)$ against $ln(Vehicles)$ together with estimated linear relationship,95% confidence interval and 95% prediction interval are displayed in figure 1, on the other hand, the plot with $PM10$ against $Vehicles$ together with estimated linear relationship,95% confidence interval and 95% prediction interval are displayed in figure 2. In both figures,the blue ribbon represented the confidence interval while the red ribbon represented the prediction interval.

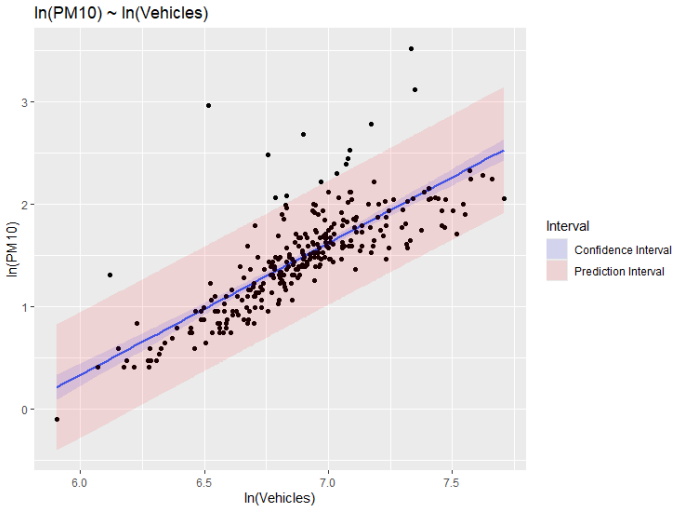


Figure 1: $ln(PM10)$ against $ln(Vehicles)$ points,fitted line and intervals

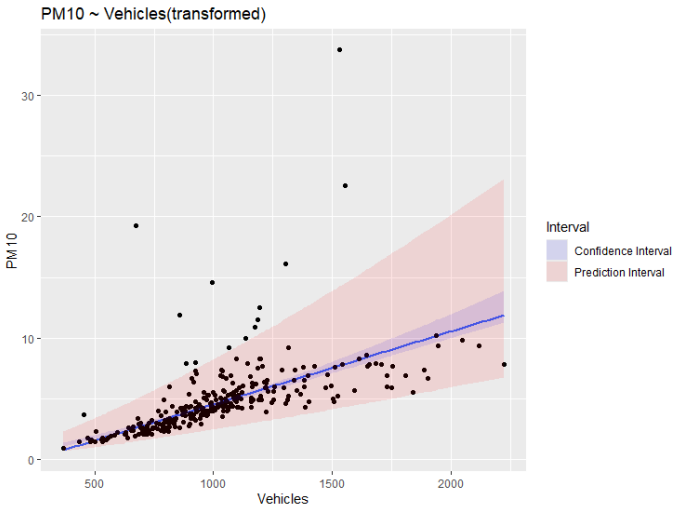


Figure 2: $PM10$ against $Vehicles$ points, fitted line and intervals

From the figures above, it is clear that the confidence interval in figure 2 expands exponentially as the number of vehicles gets larger, indicating larger variances. This observation may suggest there may be an increasing residual as the prediction of the model gets inaccurate over the span of the x-axis, a.k.a the number of the vehicles.

1.(c)

To obtain the 95 % confidence interval, the 95 % confidence interval of the β value is utilized in the calculation. The decreased PM10 particles can be calculated via $change_PM10 = (\exp(ln(0.9) * \beta_1) - 1) * 100$, while the reduced cars can be calculated via: $reduction = 0.5^{\beta_1} * 100$.

After calculation in R, the expected emission of PM10 particles would decrease 12.70832 % if the number of vehicles decrease by 10%(95 % confidence interval: [-13.74852, -11.59846]).

Additionally, in order to half its PM10 emissions,a municipality would have to reduce 40.982 % of the cars(95 % confidence interval: [37.79376, 44.43952]).

2 Part 2(a)

To test if there is a significant linear relationship between log(PM10) and log(Vehicles), we're using t-test here.

```
> PM_veh_sum <- summary(PM_veh)
> cbind(PM_veh_sum$coefficients, confint(PM_veh)|>round(digits = 2))
                Estimate Std. Error   t value     Pr(>|t|) 2.5 % 97.5 %
(Intercept)   -7.389116 0.40918648 -18.05807 2.849231e-49 -8.19  -6.58
log(Vehicles)  1.286933 0.05936672  21.67768 1.825454e-62  1.17   1.40
```

From the figure above, we can see that $|t| = 21.68$, P-value ≈ 0 <0.05 and confidence interval doesn't cover 0. Thus we should reject that $H_0 : \beta_1 = 0$, so log(Vehicles) does have significant effect on log(PM10).

2.(b)

After turning the two categorical variables Part and Coastal into factor variables, here a table of the number of all 6 combinations. And the reference category here is "Coastal: No/Part: Gotaland", and it does make sense since it has the largest number of observations, which is the rule of thumb that we want to follow.

Coastal	Part	No. of Observations
No	Gotaland	92
No	Svealand	74
No	Norrland	37
Yes	Gotaland	48
Yes	Svealand	22
Yes	Norrland	17

Table 2: no. of observations of 6 combinations

And here's the beta coefficients for the new model(added Part/Coastal interaction):

$$ln(PM10) = \beta_0 + \beta_1 ln(x_1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_2 x_3 + \beta_6 x_3 x_4 + \epsilon \quad (1)$$

with x_1 = Vehicles, x_2 = Coastal, x_3= Part: svealand, x_4= Part: Norrland.

Variable	Estimate	95% C.I.
Intercept	-8.56	(-9.59, -7.52)
log(Vehicles)	1.46	(1.31, 1.61)
Coastal: Yes	0.04	(-0.07, 0.14)
Part: Svealand	-0.09	(-0.18, -0.001)
Part: Norrland	-0.26	(-0.40, -0.12)
Coastal: Yes/Part: Svealand	0.12	(-0.06, 0.296)
Coastal: Yes/Part: Norrland	0.35	(0.14, 0.55)

Table 3: no. of observations of 6 combinations

To test if any of the added β-parameters are significantly different from zero, we'll be using the partial F test since not all parameters are tested. With the null hypothesis that, $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ then H_1 : at least one of $\beta_2, \beta_3, \beta_4 \neq 0$. We will leave the interaction terms for now, as we'll deal with it in the next step.

Thus, we applied the analysis of variance between *Model.1(b)* and the following model:

$$ln(PM10) = \beta_0 + \beta_1 ln(x_1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \tag{2}$$

Then we have the results with test statistics $F = 4.8496$. If the null hypothesis stands, then F should be F(p, n - (p + 1)) = F(3, 285) distributed. And the P value of 0.002627 <0.05, which means we can reject H_0 for $\alpha = 0.05$ so that we can conclude that at least one of the new parameters isn't zero.

The same applies for the interaction terms. We'll be using the analysis of variance between *Model.1(b)* and the model without the interaction terms. The results came in with test statistics $F = 5.6732$, again if H_0 holds, F should be F(p, n - (p + 1)) = F(2, 283) distributed. And the P value of 0.003839 <0.05. Therefore, at one of the interaction terms isn't zero.

Below are the tables for the estimates and 95% confidence interval for all 6 possible combinations of both expected logPM10 and PM10 respectively, given that when Vehicles=1000 per 1000 inhabitants.

	Estimates	95% C.I
Yes/Gotaland	1.580244	(1.493028, 1.667459)
Yes/Svealand	1.606324	(1.47017, 1.742478)
Yes/Norrland	1.664842	(1.522806, 1.806877)
No/Gotaland	1.543648	(1.482476, 1.604821)
No/Svealand	1.451213	(1.382906, 1.519519)
No/Norrland	1.281529	(1.165516, 1.397543)

Table 4: Expected log-PM10 and 95% confidence intervals for *Model.2b*

	Estimates	95% C.I
Yes/Gotaland	4.856139	(4.450551, 5.298688)
Yes/Svealand	4.984455	(4.349975, 5.711479)
Yes/Norrland	5.284836	(4.585073, 6.091395)
No/Gotaland	4.68164	(4.403835, 4.97697)
No/Svealand	4.268288	(3.986471, 4.570028)
No/Norrland	3.602144	(3.207577, 4.045248)

Table 5: Expected PM10 and 95% confidence intervals for *Model.2b*

2.(c)

To further reduce to the number of necessary combinations, we must first observe the similarities in the model. From table 5, we can see that the estimates and C.I. from (Coastal, Parts): (Yes,Gotaland), (Yes,svealand), (No,Gotaland) and (No,svealand) are quite similar in a sense. Thus we create a new variable based on this.

$$\begin{aligned} &\text{New Variable} \\ &\text{Value 1: Part = Gotaland | svealand} \\ &\text{Value 2: Part = Norrland \& Coastal = Yes} \\ &\text{Value 3: Part = Norrland \& Coastal = No} \end{aligned} \tag{3}$$

Then, using this reduced version we have its estimates and confidence intervals in the table below.

	Estimates	95% C.I
Yes/Gotaland	4.587944	(4.404126, 4.779434)
Yes/Svealand	4.587944	(4.404126, 4.779434)
No/Gotaland	4.587944	(4.404126, 4.779434)
No/svealand	4.587944	(4.404126, 4.779434)
Yes/Norrland	5.306225	(4.598489, 6.122886)
No/Norrland	3.672487	(3.274647, 4.118661)

Table 6: Expected PM10 and 95% confidence intervals for proposed model

However, if we inspect even more carefully, the estimate for Part = Svealand and Coastal = No actually lies outside of the confidence interval when compared to table 5. We take this as a sign that two models still have quite some differences, so we propose another variable which narrows down the condition of value 1 to further fit the target model.

$$\begin{aligned} &\text{New Variable} \\ &\text{Value 1: Part = Gotaland | Part = svealand \& Coastal = Yes} \\ &\text{Value 2: Part = Svealand \& Coastal = No} \\ &\text{Value 3: Part = Norrland \& Coastal = Yes} \\ &\text{Value 4: Part = Norrland \& Coastal = No} \end{aligned} \tag{4}$$

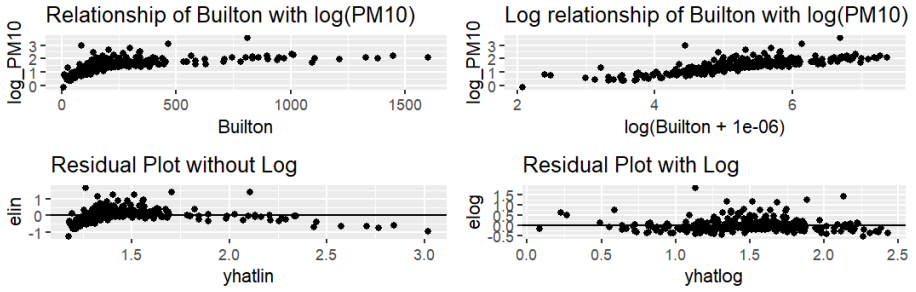
With this, we have its estimates and confidence intervals in table 7 below.

	Estimates	95% C.I
Yes/Gotaland	4.759556	(4.532658, 4.997812)
Yes/Svealand	4.759556	(4.532658, 4.997812)
No/Gotaland	4.759556	(4.532658, 4.997812)
No/svealand	4.262583	(3.98191, 4.563039)
Yes/Norrland	5.296068	(4.596416, 6.102219)
No/Norrland	3.638948	(3.247902, 4.077076)

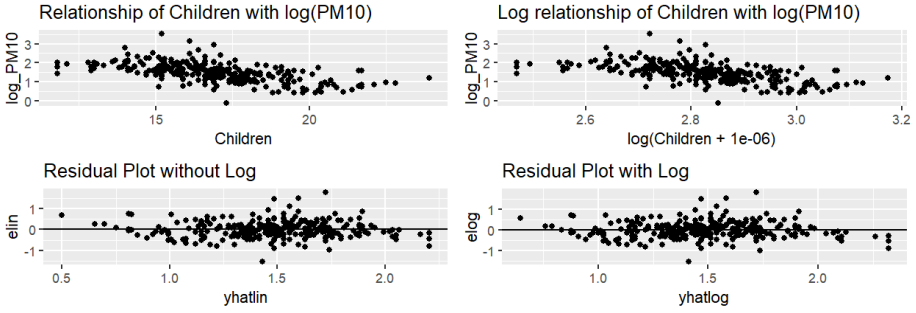
Table 7: Expected PM10 and 95% confidence intervals for *Model.2(c)*

2.(d)

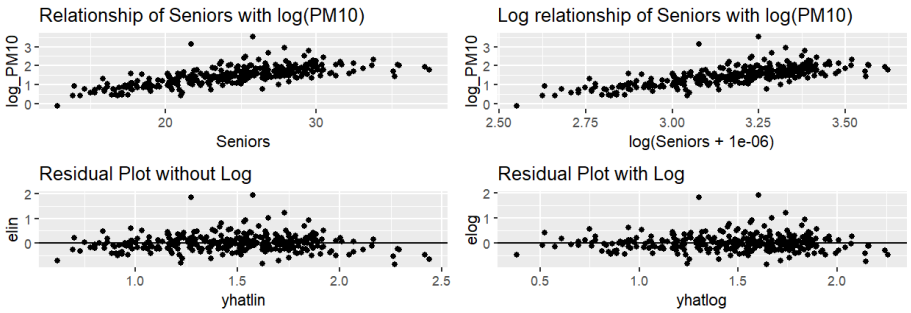
In order to decide which variable should be log-transformed and which are the two most stronglyed correlated, we begin by plotting the residual and prediction plots for variable Builton, Children, Senior, Higheds, Income and GRP.



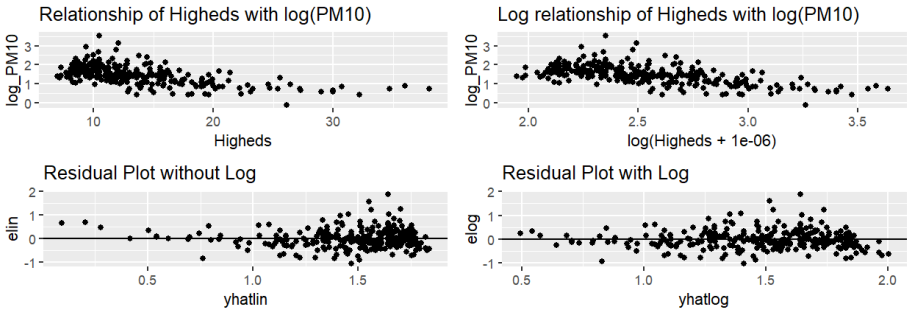
Plots for model with variable Builton



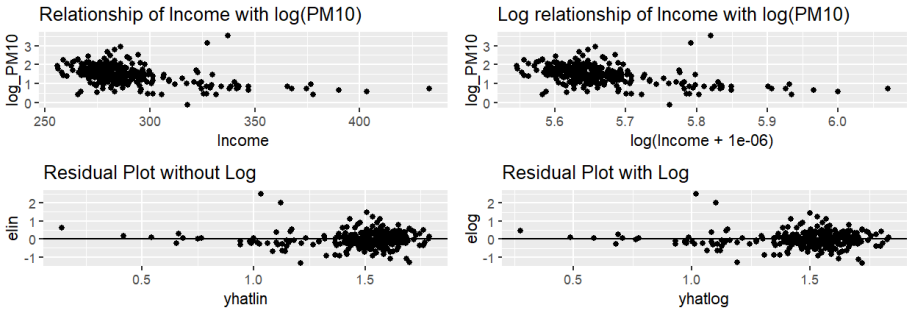
Plots for model with variable Children



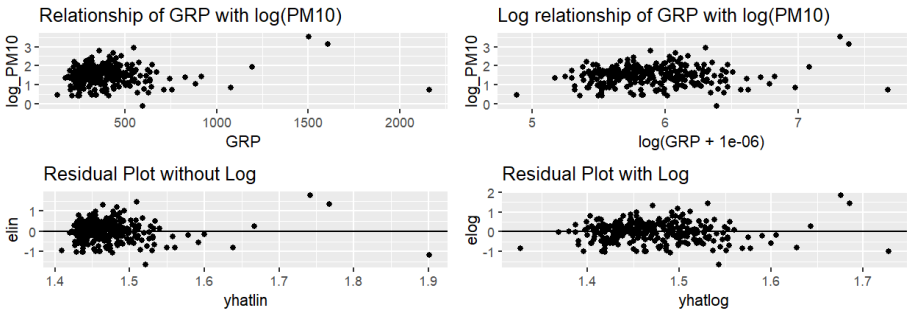
Plots for model with variable Senior



Plots for model with variable Higheds



Plots for model with variable Income



Plots for model with variable GRP

As the plots above shown, we decided that all variables should be log-transformed except for "Children" since the distributions have become less skewed after the transformation and the residuals are gathering more on the x=0 axis.

Also we can see that "log(Builton)" and "log(Seniors)" are the two that are most strongly correalted to log(PM10). Base on this, we have the model:

$$ln(PM10) = \beta_0 + \beta_1 ln(\text{Vehicles}) + \beta_2 ln(\text{Seniors}) + \beta_3 ln(\text{Builton}) + \epsilon \qquad (5)$$

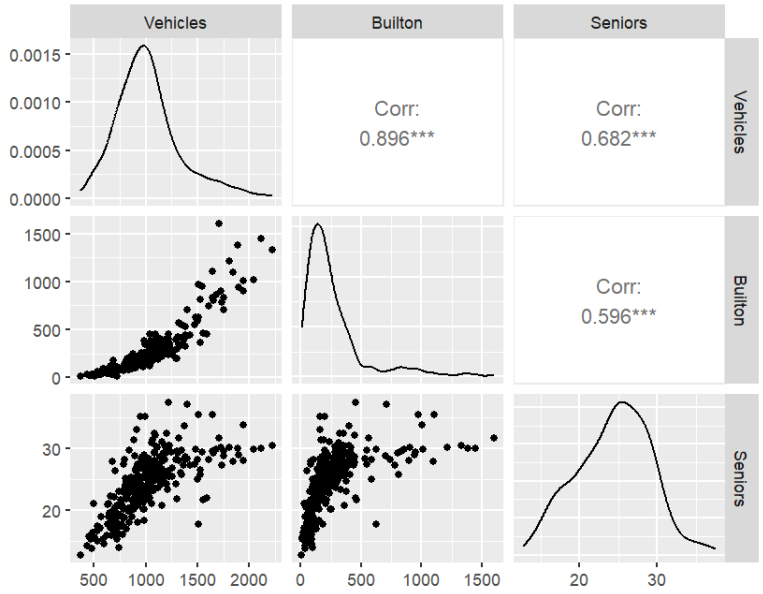
And its estimates, standard errors, P-values and confidence intervals:

	Estimates	Std. err	P	C.I.
β_0	-5.6554913	0.87353331	4.135979e-10	(-7.374861010, -3.9361216)
β_1	0.6532632	0.16319072	7.973372e-05	(0.332056041, 0.9744704)
β_2	0.6405519	0.13275468	2.279905e-06	(0.379251807, 0.9018521)
β_3	0.1142812	0.05917382	5.443657e-02	(-0.002190271, 0.2307526)

Table 8: estimates, std errors, P-values and confidence intervals for the model

To determine which variable we should discard, we used pairwise correlation plot and calculate the VIF values in the table below.

	VIF value
log(Vehicles)	8.458155
log(Seniors)	2.540938
log(Builton)	9.268623



Since log(Builton) and log(Vehicles) are highly correalted from the plot above, we choose to discard log(Builton). Eventually, we have *model.2(d)*:

$$ln(PM10) = \beta_0 + \beta_1 ln(\text{Vehicles}) + \beta_2 ln(\text{Seniors}) + \epsilon \tag{6}$$

	Estimates	Std. err	P	C.I.
β_0	-7.1662900	0.39056044	2.720485e-50	(-7.9350162, -6.397564)
β_1	0.9223340	0.08537629	4.718633e-23	(0.7542909, 1.090377)
β_2	0.7206029	0.12671629	3.186581e-08	(0.4711917, 0.970014)

Table 9: estimates, std errors, P-values and confidence intervals for the model

And the updated VIF values.

	VIF value
log(Vehicles)	2.293231
log(Seniors)	2.293231

In conclusion, the updated VIF values are lower than the original ones, which means there's less dependence on the other variables.

Part 3. MOdel validation and selection

3.(a) Leverage

The leverage from Model 2e against the linear predictor are shown the the figure 3.

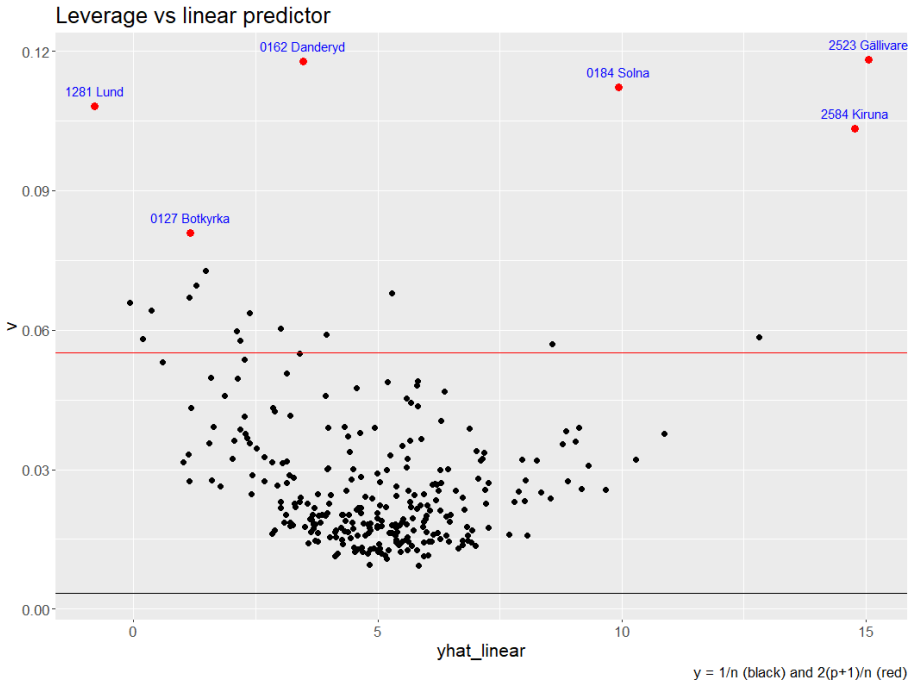


Figure 3: Leverage vs. Log predictor

After eliminating the x-variables which are highly correlated to each other, we are left with the x-variables in turn to be $[log(Vehicles), Children, log(Higheds), log(Income), log(GRP)]$. The pairs of x-variables against each other and separately for each category are plotted below.

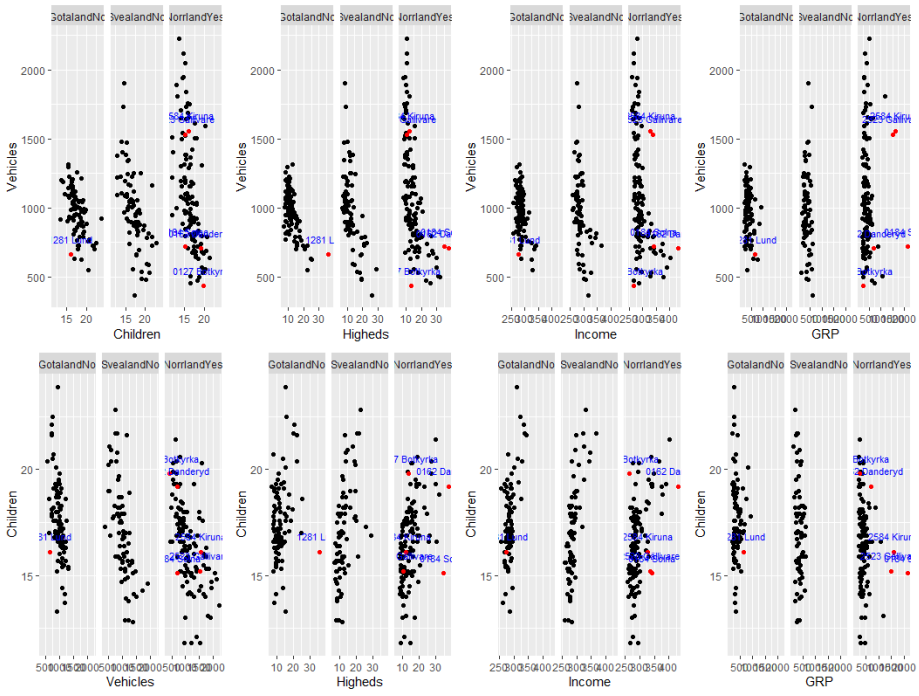


Figure 4: Vehicles,Children against other variables

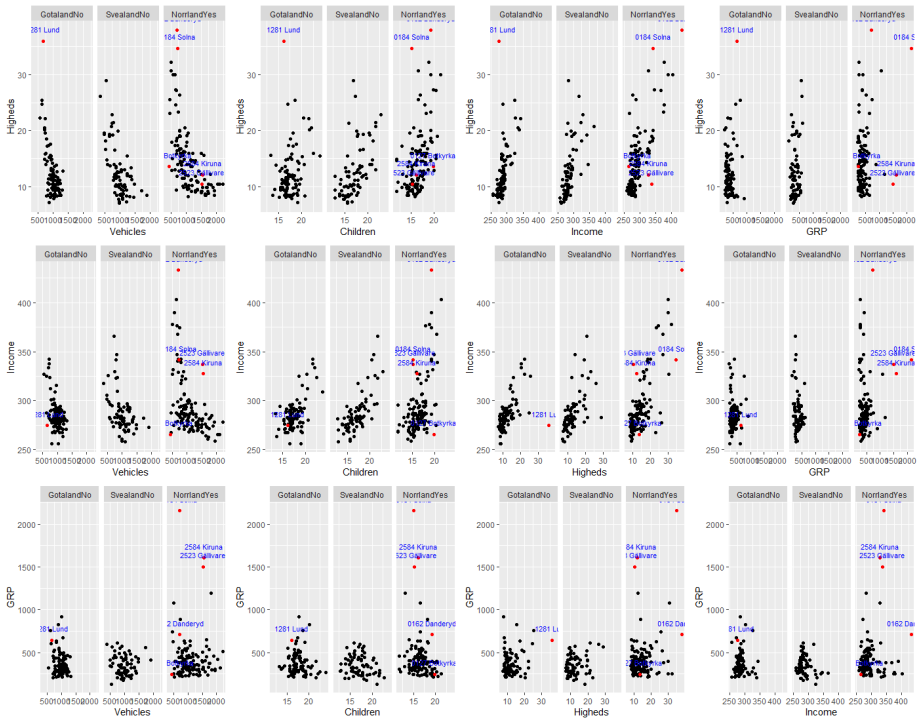


Figure 5: Higheds,Income,GRP against other variables

From figures above,it can be seen that the GRP, Income, and Higheds remain prominent against other variables, which suggests they contributed to the leverage more than other variables.

3.(b) Cook's distance

In this section, the Cook's distance of Model 2e with 6 municipalities having the highest cook's distance shall be displayed in the plot. From figure 6,it could be observed that the 6 municipalities are not identical to which had high leverage.DFBETAS suggests the municipalities that have the maximum absolute DFBETAS values are in turn to be "1480 Göteborg","2584 Kiruna","1761 Hammarö", "1761 Hammarö","2523 Gällivare","2523 Gällivare","1761 Hammarö","0481 Oxelösund".

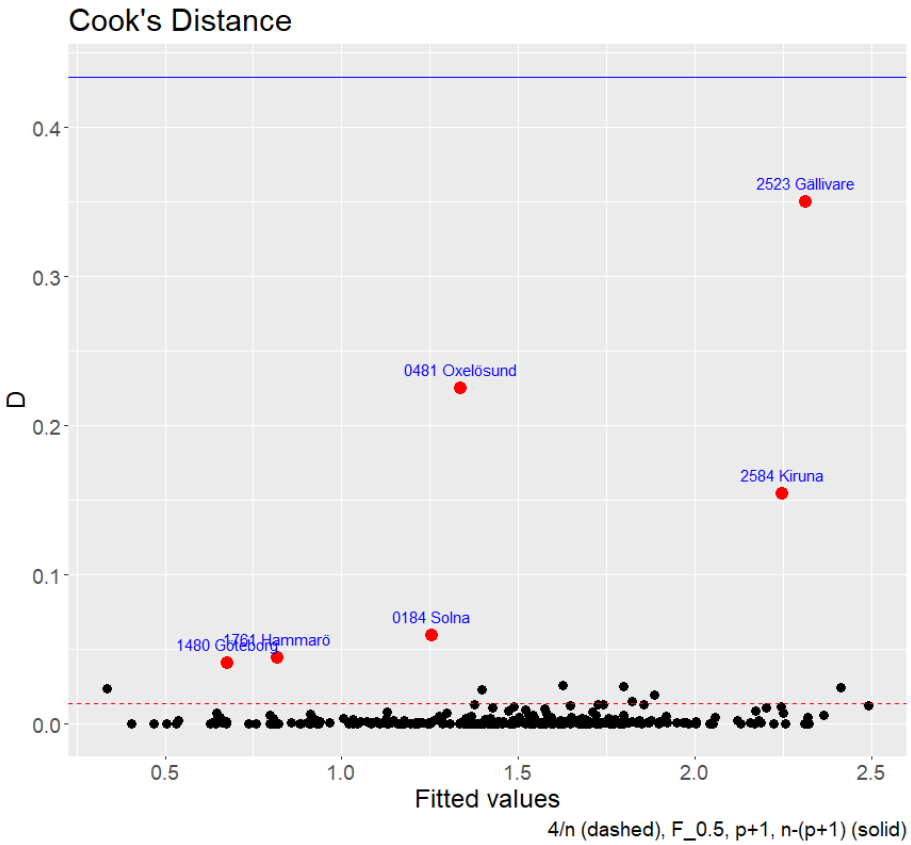


Figure 6: Cook's Distance

Additionally, compared the DFBETAS with $2/\sqrt{n}$, we have filtered the beta values that are the most impacted by observation, which are $[log(Vehicles), Children, log(Higheds), log(Incom$ to see the reason for the municipalities above to have large influence on the parameters, we plotted the log(PM10) against independent variables in figure 7 to figure 11, from which we could conclude that the influential municipalities are outliers in all of the plots and thus have a great impact on it.

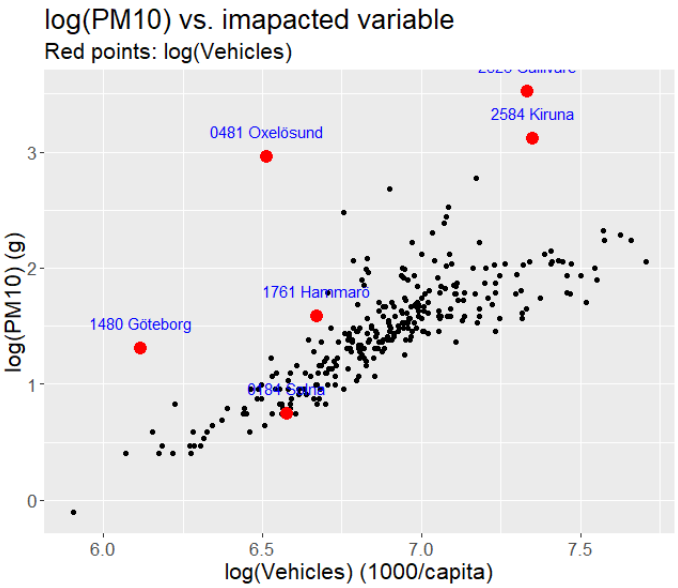


Figure 7: log(PM10) log(Vehicles)

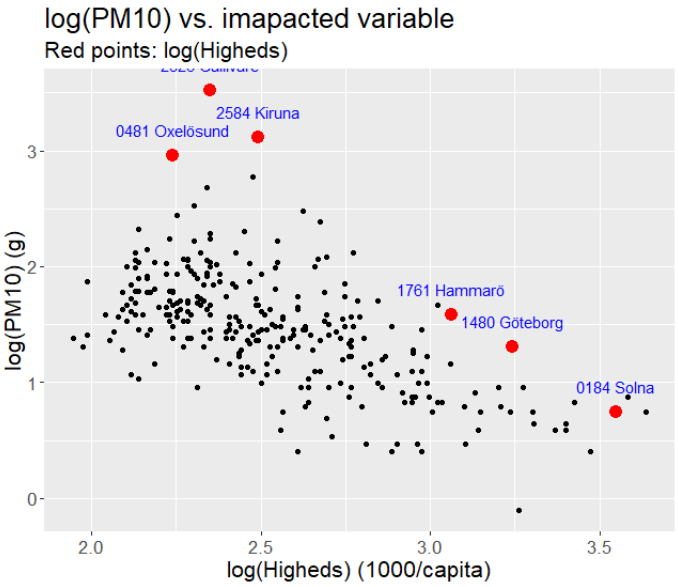


Figure 8: log(PM10) log(Higheds)

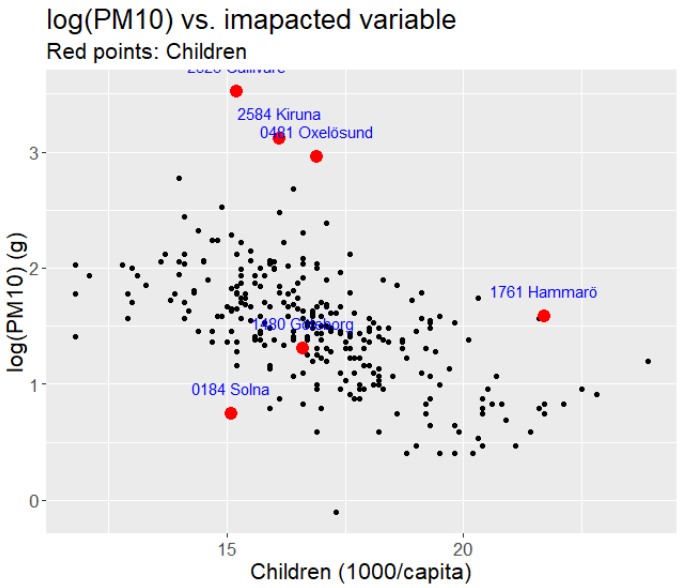


Figure 9: log(PM10) log(Vehicles)

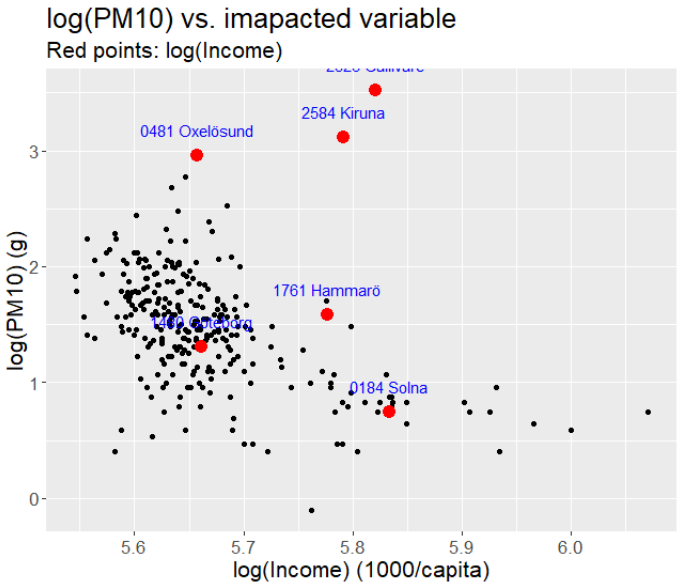


Figure 10: log(PM10) log(Vehicles)

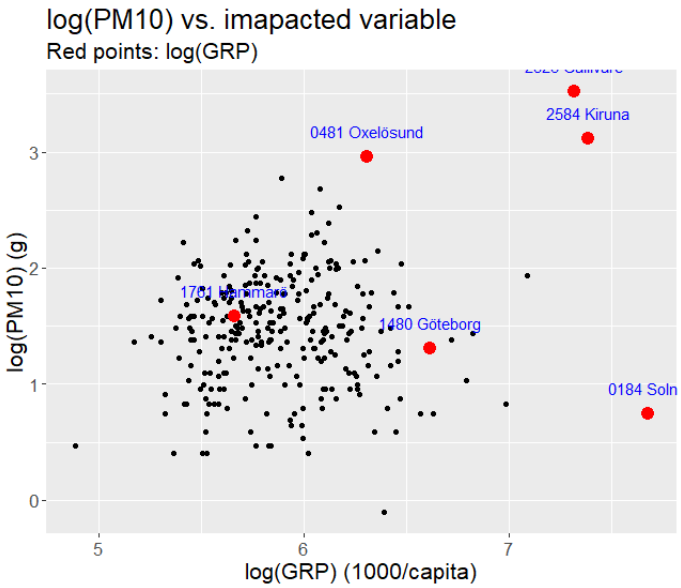


Figure 11: log(PM10) log(Vehicles)

3.(c) studentized residuals

In this section, studentized residuals r_i^* and squared absolute studentized residuals against linear predictor are asked to be plotted.

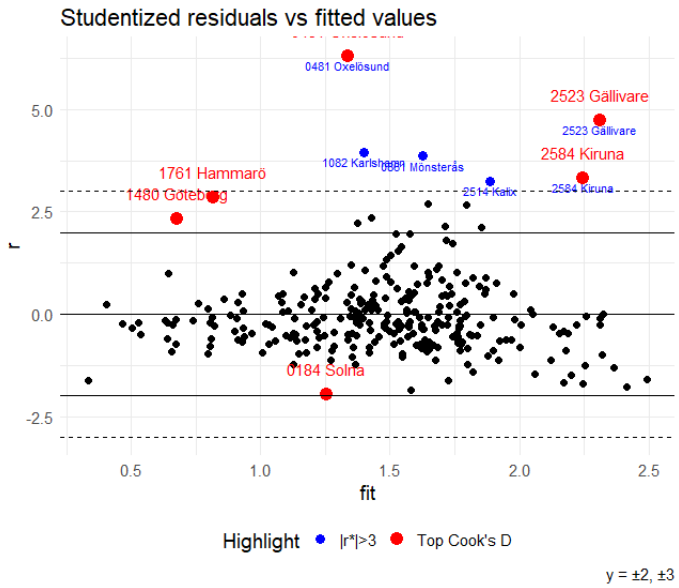


Figure 12: Studentized residuals against fitted values

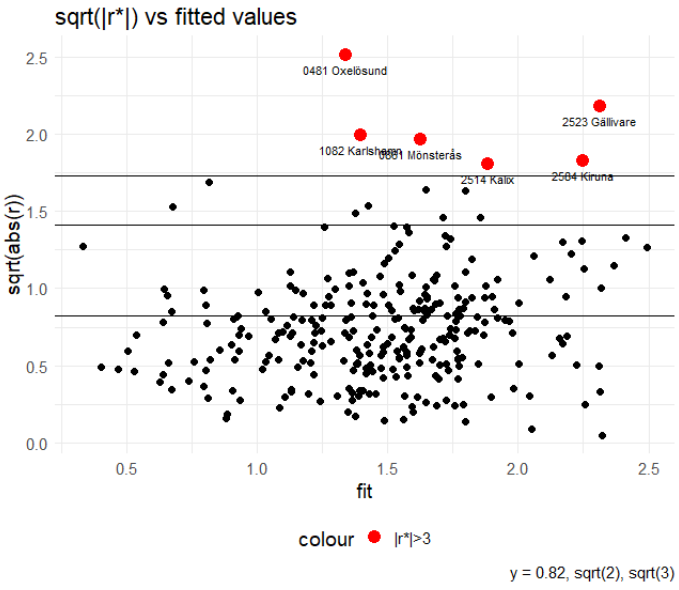


Figure 13: Sqrt absolute studentized residuals vs. fitted values

From the above figure 12, the red points represented all municipalities with the largest Cook's distance have large residuals. While the municipalities which

don't have large Cook's distance yet fulfilled the condition $|r_i^*| > 3$ are: 0861 Mönsterås,1082 Karlshamn,2514 Kalix. Besides, in figure 13, no systematic patterns in the residuals, and are randomly distributed around the median,0,82. All of the municipalities with large Cook's distance lie above the reference line $y = \sqrt{3}$.

3.(d) Explain,exclude,refit

First, in this task, after five rounds of exculsions on the municipalities,all the removed municipalities are listed below in table 10:

Municipality	Emission Source or Company name
0481 Oxelösund	SMA Mineral AB SAAB
2584 Kiruna	iron etc.
1082 Karlshamn	Soedra Cell
2523 Gällivare	iron etc.
0861 Mönsterås	Soedra Cell
2514 Kalix	Billerud AB
1882 Askersund	Ahlstrom
1484 Lysekil	Preeem AB
1761 Hammarö	paper mill
1480 Göteborg	Preem AB,Stl Refinery AB
1494 Lidköping	Swedish Air Force wing
1471 Götene	dairy
2262 Timrå	paper mill
1781 Kristinehamn	Nordic paper
1460 Bengtsfors	Ahlstrom
1885 Lindesberg	Billerud AB
0980 Gotland	factory, quicklim
0319 Alvkarleby	Stora Enso
1272 Bromölla	- paper mill

Table 10: Removed Municipalities

After removing the problematic municipalities,we refitted the model and compared it with the old one, their beta values and the confidence intervals are listed below in the table 11 and table 12.

Variables	β	95% CI
Intercept	-5.349	[-8.797,-1.901]
log(Vehicles)	0.909	[0.751,1.068]
log(Higheds)	-0.365	[-0.527,-0.204]
log(Income)	0.203	[-0.420,0.825]
Children	-0.0377	[-0.06,-0.01]
log(GRP)	0.171	[0.076,0.266]

Table 11: β estimates of the old model 2e

Variables	β	95% CI
Intercept	-2.018	[-4.262,0.227]
log(Vehicles)	1.0	[0.905,1.203]
log(Higheds)	-0.149	[-0.252,-0.046]
log(Income)	-0.435	[-0.83,-0.04]
Children	-0.029	[-0.0434, -0.016]
log(GRP)	-0.01	[-0.07,0.05]

Table 12: β estimates of the new model 3d

In the new model the log(GRP) and INtercept are not significant anymore while in the old model, log(Income) is not significant. Since log(GRP) was mot discovered as having the highest DFBETA in 3(b), this is maybe because high residual obeservation has strong impact on the β value of log(GRP). Additionally, the log(Income) and log(GRP) may correlate. Finally, the normality and the homoscedacity of the new model can be verified via Q-Q plot and Square root of the absolute studentized residuals plot respectively.

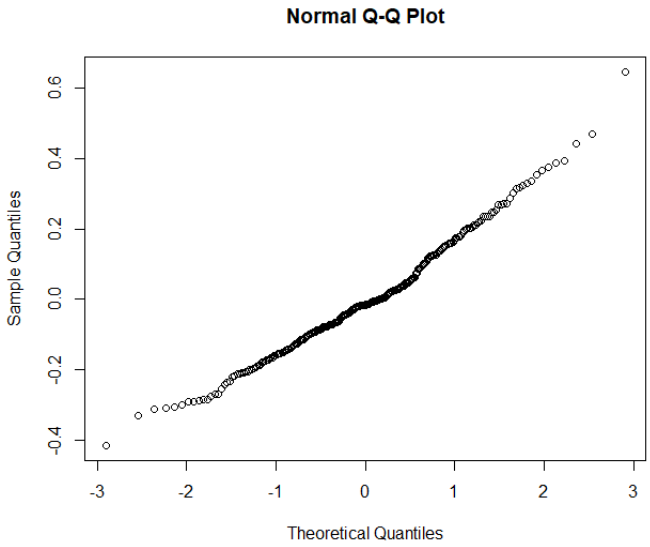


Figure 14: Q-Q plot of the new model

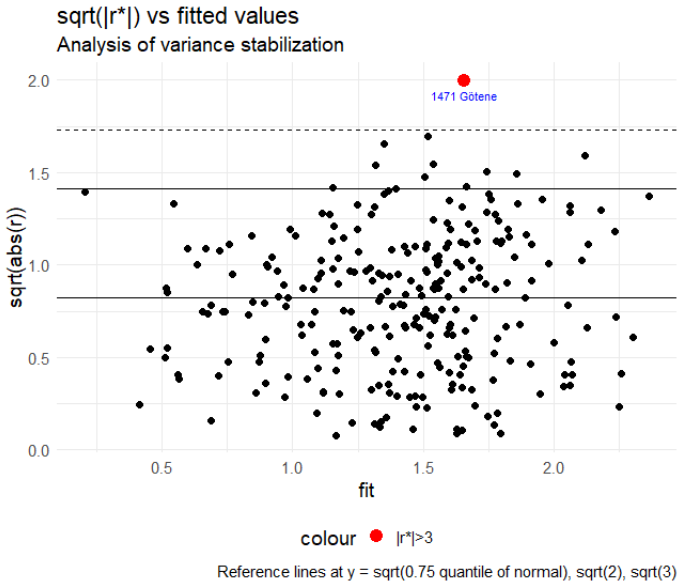


Figure 15: Square root of the absolute studentized residuals

3.(e) Variable selection

In this section,the clean data are used to select the best model, in the table below 13, all the measures of the models are listed for model selection.

model	β	res$_S D$	R^2	Adjusted R^2	AIC	BIC
Null	1	0.436	0	0	323.4	330.6
1b	2	0.194	0.802	0.801	-114	-103.5
2c	3	0.187	0.826	0.816	-133.5	-115.5
3d	4	0.167	0.856	0.852	-189.03	-156.6
AIC	5	0.167	0.856	0.853	-190.9	-162.1
BIC	6	0.168	0.854	0.851	-188.23	-163.01

Table 13: Comparison between models

To asses how much the variability of log(PM10) can be explained,the adjusted R^2 would do. Besides, we also look