Lund University

FMSN30 / MASM22 / FMSN40 - Linear and Logistic Regression
Spring 2024

Project 1: Linear Regression

Group P1 10

Elise Olsson[1], Yifan Zhang[2]

[1] el6373ol-s@student.lu.se
[2] yi6136zh-s@student.lu.se

Introduction

In this project we want to find a model to model the yearly emissions of atmospheric particles with a diameter between 2.5 and 10 μm, PM_{10}, per capita in the 290 municipalities in Sweden. The model is based on data from Statistics Sweden. We want to find a model that models the emissions as accurate as possible while only using the necessary variables. This is done by starting out easy with one variable and then adding on more for a multiple linear regression model. Different models are then evaluated to find the best one to model the emissions.

Results

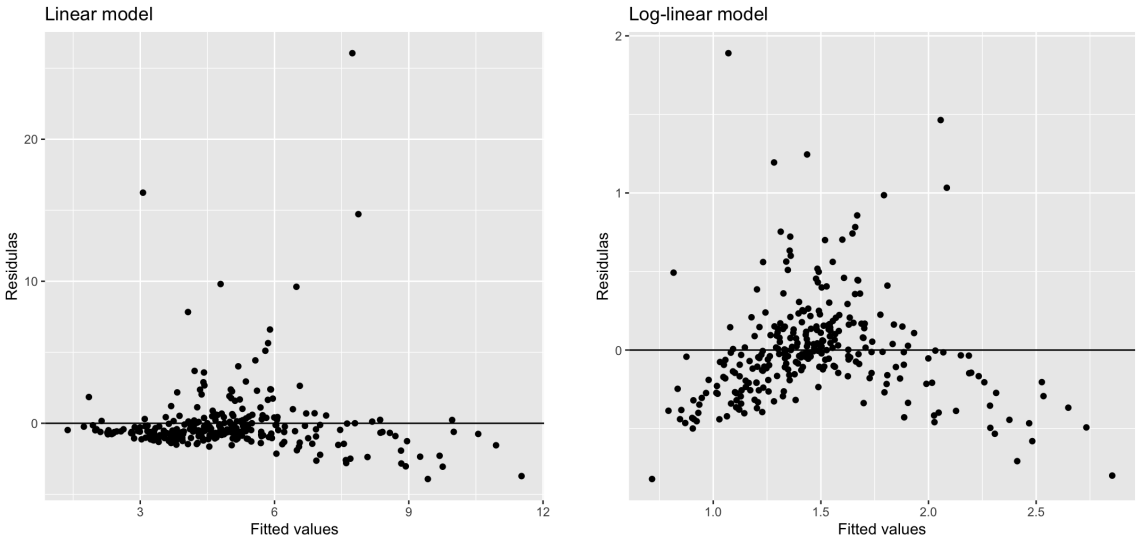
Part 1: Linear relationship between PM_{10} and vehicles

a

For starters a model based on the amount of 1000 vehicles per capita was created. To model the PM_{10} particles with vehicles as an explanatory variable the residuals were studied to see if the logarithm should be used on any of the variables. The first model

$$\ln(PM_{10}) = \beta_0 + \beta_1 x + \epsilon$$

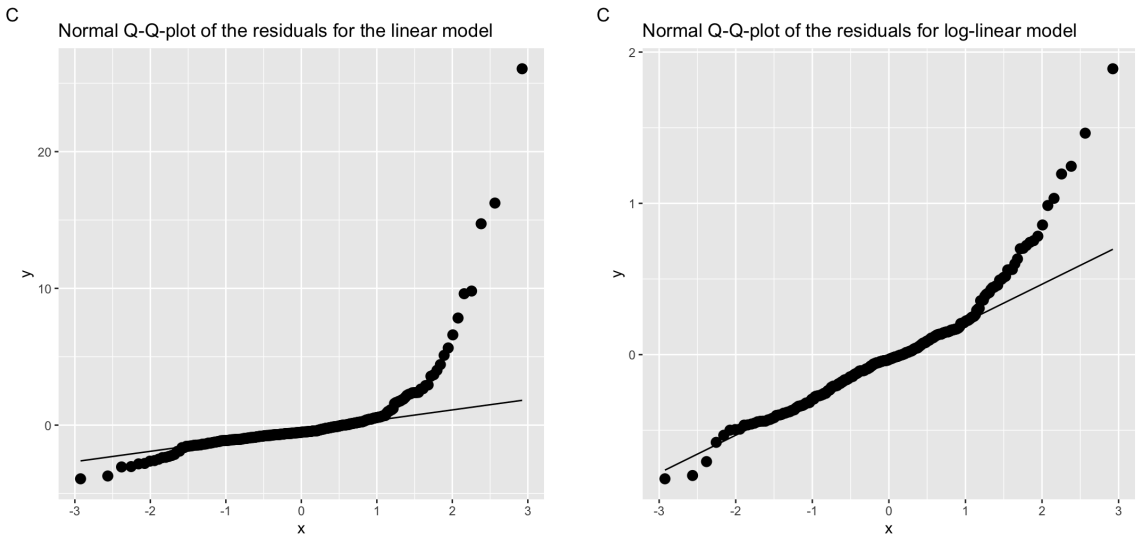
with x = Vehicles, was compared to the same model but without the logarithm. The residuals were plotted both against the fitted values and in Q-Q-plots, see figures 1a, 1b, 2a and 2b.



(a) Linear model.

(b) Log-linear model.

Figure 1: The residuals for the linear model and the log-lin model against the fitted values.



(a) Linear model.

(b) Log-linear model.

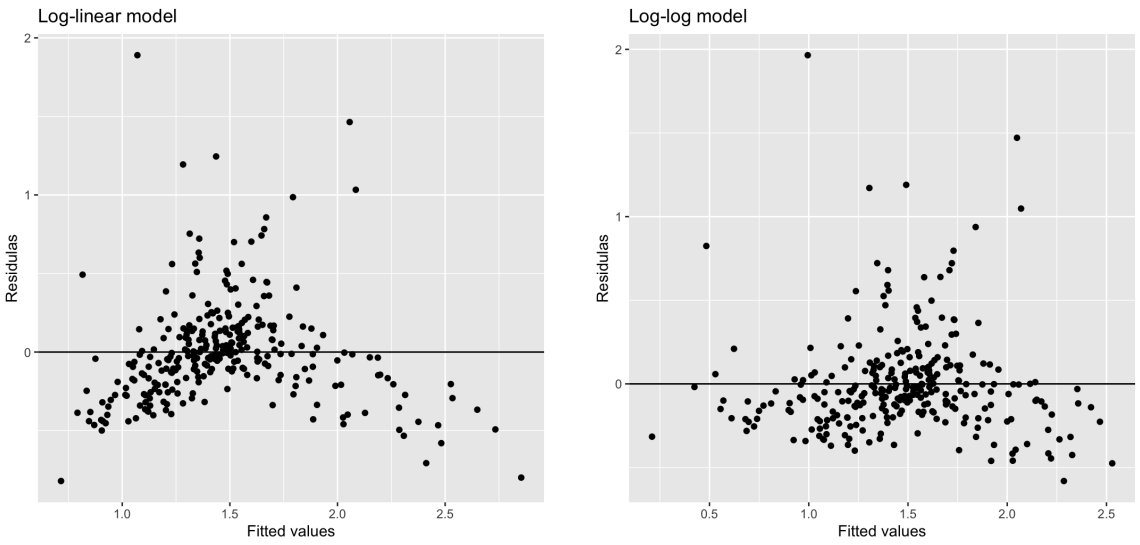
Figure 2: The Q-Q-plot for the linear model and the log-lin model.

The logarithmic-linear model has much smaller residuals than the linear model as seen in figures 1a and 1b, the log-lin model also follows the straight line in the Q-Q-plot much better than the linear model. The linear model has many high residuals and many quantiles outside the line in the Q-Q-plots. This suggests that the log-lin model is the better choice so that one is picked.

The same two plots were also done for the following model

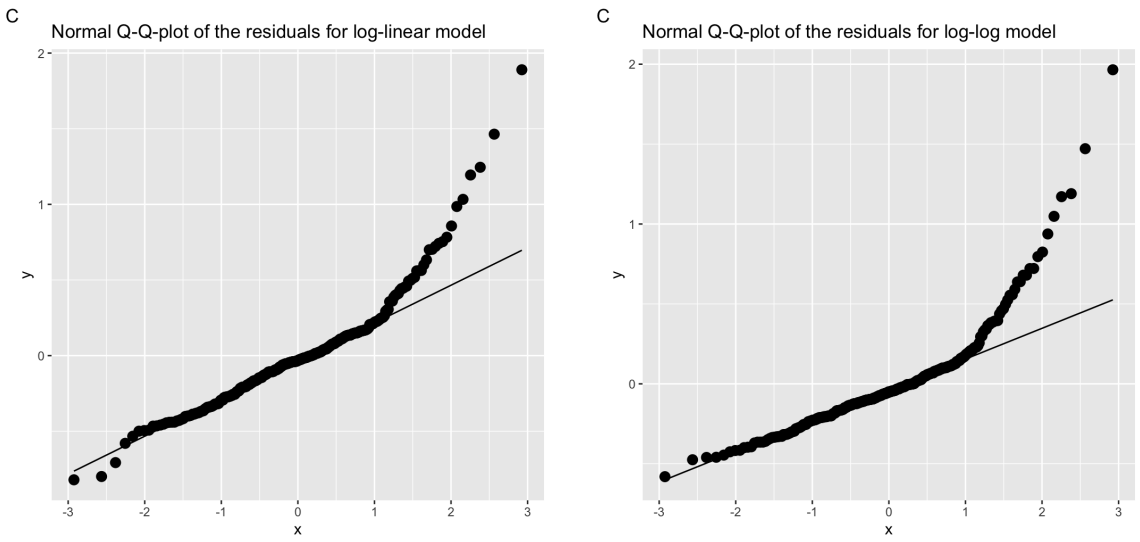
$$\ln(\mathrm{PM}_{10}) = \beta_0 + \beta_1 \ln(x) + \epsilon$$

to see if $x = \ln(\text{Vehicles})$ or $x = \text{Vehicles}$ is better to use for our model. The residuals and Q-Q-plot for the log-log model can be seen in figure 3b and 4b and those were compared to with the plots for the log-lin model in figure 3a and 4a.



(a) Log-linear model. (b) Log-log model.

Figure 3: The residuals for the log-lin model and the log-log model against the fitted values.



(a) Log-lin model. (b) Log-log model.

Figure 4: The Q-Q-plot for the log-lin model and the log-log model.

The Q-Q-plot seems to fit better for the log-log model and the residuals are also a little bit smaller for that model compared to the log-lin model. x is therefore set to $x = \ln(\text{Vehicles})$ and our model (*Model.1(b)*) is

$$\ln(\mathrm{PM}_{10}) = \beta_0 + \beta_1 \ln(\mathrm{Vehicles}) + \epsilon.$$

b

The β estimates were calculated for *Model.1(b)* and their estimates and 95 % confidence intervals can be seen in table 1.

Table 1: The β-estimates and confidence intervals for *Model.1(b)*.

β-coefficient	Estimate	Lower interval	Upper interval
β_0	-7.389	-8.194	-6.587
β_1	1.286	1.170	1.404

The $\ln(\mathrm{PM}_{10})$ values were plotted against ln(Vehicles) together with the confidence interval and prediction interval in figure 5.

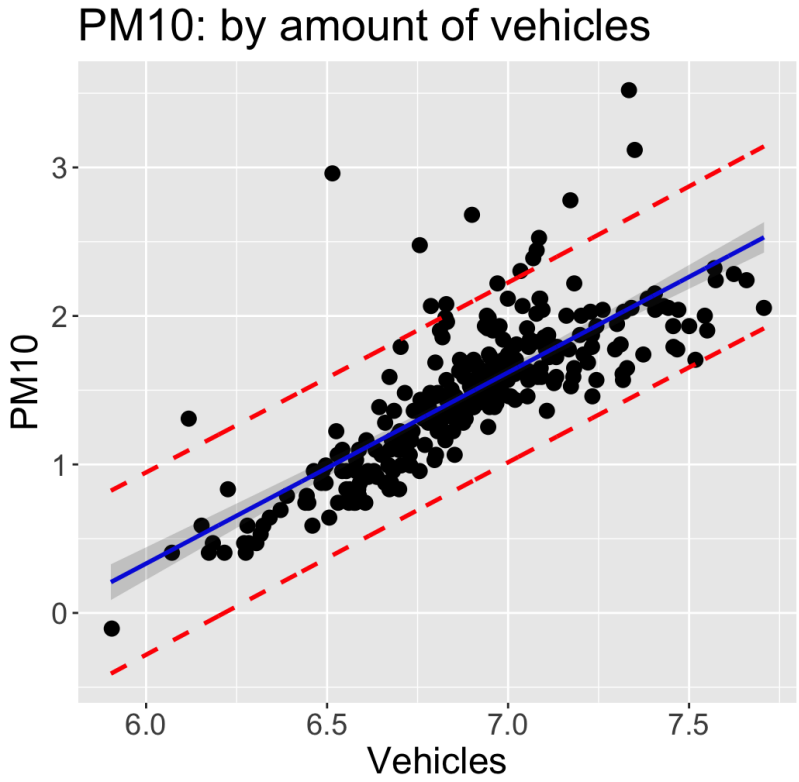


Figure 5: The fit of the log-log model together with the 95 % confidence (in blue) and prediction intervals (in red).

The relationship between PM_{10} and vehicles was also transformed back from the logarithmic

relationship to PM_{10} = ..., and the same plot with the confidence and prediction intervals was done for the transformed relationship, see figure 6.

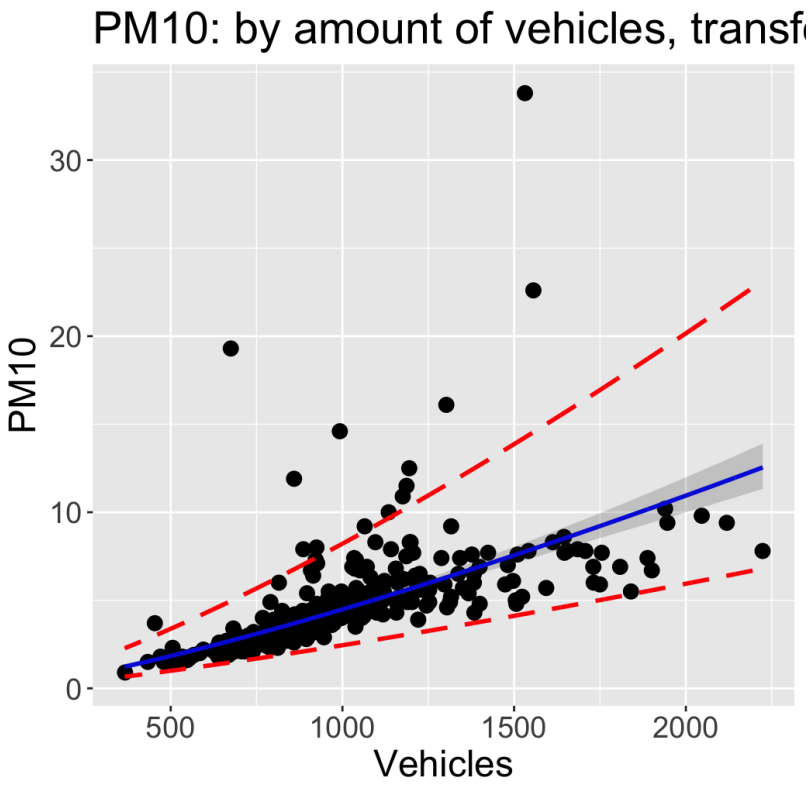


Figure 6: The fit of the transformed log-log model together with the 95 % confidence (in blue) and prediction intervals (in red).

The model seems to fit well to the data except for a few outliers that are outside the prediction interval. The data points outside the intervals and far away from the fitted line could be affecting the model and making the fit worse and those points are the same points with large residuals.

c

If the number of vehicles would decrease by 10% the change in emissions of $\ln(PM_{10})$ particles would change by $\beta_1 \cdot \ln(0.9)$, with $\beta_1 = 1.286$ this equals to -0.1356. For PM_{10} this 10 % decrease in vehicles would mean a 12.68% decrease in PM_{10} emissions, with confidence interval $[-13.84\%, -11.52\%]$.

$$\ln(\text{Vehicles}_{\text{new}}) = \ln(0.9 \times \text{Vehicles}) = \ln(0.9) + \ln(\text{Vehicles})$$
$$\ln(PM_{10,\text{ new}}) = \beta_0 + \beta_2 \times (\ln(\text{Vehicles}) - 0.10536)$$
$$\ln(PM_{10,\text{ new}}) = \beta_0 + \beta_2 \times \ln(\text{Vehicles}) - \beta_2 \times 0.10536 \frac{PM_{10,\text{new}}}{PM_{10}} = e^{-\beta_2 \times -0.1356}$$
$$1 - e^{\Delta ln(PM10) \approx e^{-0.1356}} \approx 0.1268$$

For a municipality to half its PM_{10} emissions the number of cars would have to be reduced by 58.35 %, with a confidence interval of $[-61.02\%, -55.31\%]$.

$$\begin{aligned} \ln(PM_{10\text{new}}) &= \ln(0.5 \times PM_{10\text{n}}) = \ln(0.5) + \ln(PM_{10}) \\ \ln(PM_{10}) - \ln(0.5) &= \beta_0 + \beta_1 \ln(Vehicles_{\text{new}}) \\ \ln(PM_{10}) &= \beta_0 + \beta_1 \ln(Vehicles) \end{aligned}$$

After deriving the expression for $Vehicles_{\text{new}}$, we can calculate the percentage reduction in $Vehicles$ required when PM_{10} decreases by 50%.

$$\begin{aligned} \beta_0 + \beta_1 \ln(Vehicles) - \ln(0.5) &= \beta_0 + \beta_1 \ln(Vehicles_{\text{new}}) \\ \ln(Vehicles_{\text{new}}) &= \ln(Vehicles) - \frac{\ln 0.5}{\beta_1} \\ Vehicles_{\text{new}} &= Vehicles \times e^{-\frac{\ln 0.5}{\beta_1}} \approx Vehicles \times 0.5835 \end{aligned}$$

Part 2: PM_{10} with other explanatory variables

a

We used a t-test to evaluate whether the regression coefficient β_1 of log(Vehicles) in *Model.1(b)* is significantly different from zero. The null hypothesis (H_0) states that the coefficient equals to zero, indicating no linear relationship between log(Vehicles) and log(PM_{10}). The alternative hypothesis (H_1) suggests the coefficient is not zero, so that there is a significant linear relationship.

The t-test results are shown in table 2. The test statistic is a t-value of 21.68 and 288 degrees of freedom under the null hypothesis. With an extremely small p-value of less than 2×10^{-16} , significantly below 0.05, so we reject the null hypothesis. This confirms a significant linear relationship between log(Vehicles) and log(PM_{10}).

Table 2: T-test Results of log(PM_{10})-log(Vehicles) model

Coefficient	Estimate	Std. Error	t value	Pr($>\|t\|$)
β_0	-7.38912	0.40919	-18.06	$< 2 \times 10^{-16***}$
β_1	1.28693	0.05937	21.68	$< 2 \times 10^{-16***}$

b

From the t-test that was done it was seen that vehicles was a significant variable and the next step was to add more explanatory variables. We started by adding the two categorical variables "Part" and "Coastal" and turned them into factor variables. The coastal variable was labeled "No" and "Yes" representing inland or coastal, and the Part category was labeled "Norrland", "Gotaland" and "Svealand" representing the three parts of Sweden. With those variables there are 6 different combinations and the number of observations in each combination is presented in table 3 below.

Both the "Coastal" and "Part" variables were added to *Model.2(b)* as well as their interaction and this new model is called *Model.2(b)*. The new model is

$$\ln(\text{PM}_{10}) = \beta_0 + \beta_1 \ln(\text{Vehicles}) + \beta_2 \cdot \text{Coastal} + \beta_3 \cdot \text{Part} + \beta_4 \cdot \text{Part} \cdot \text{Coastal} + \epsilon.$$

Table 3: The number of observations in each Part/Coastal category.

Part \ Coastal	Yes	No
Svealand	74	22
Götaland	92	48
Norrland	37	17

The new β estimates and their confidence intervals can be seen in table 4.

Table 4: The β-estimates and confidence intervals for *Model.2(b)*.

β-coefficient	Estimate	Lower interval	Upper interval
β_0	-8.518	-9.535	-7.502
β_1	1.462	1.311	1.612
β_{Yes}	-0.037	-0.142	0.068
β_{Svealand}	0.026	-0.128	0.180
β_{Norrland}	0.085	-0.084	0.253
$\beta_{\text{Yes, Svealand}}$	-0.119	-0.297	0.060
$\beta_{\text{Yes, Norrland}}$	-0.347	-0.550	-0.143

The reference category is Coastal= "No" and Part= "Gotaland". This can be seen by looking at the β parameters and since Coastal = "No" and Part = "Gotaland" are not included they must be the reference categories and the other coefficients are in reference to them. Since the reference to themselves is 0 they are included in the intercept, β_0. Part = "Gotaland" is suitable as a reference since it has the greatest amount of observations. Coastal = "No" has a lot less observations than Coastal = "Yes" so Coastal = "Yes" would have been a better reference category to use.

To test if any of the added β parameters are significantly different from zero a partial F-test was done. The null hypothesis H_0 is that all added parameters equal zero and H_1 is that at least one of the added β parameters are significantly different from zero. The test statistic was 5.27 and the P-value 0.0001194, which is below 0.05, so the null hypothesis was rejected. If H_0 was true the distribution would be F(5, 290-(5+1)).

A test was also done to see if the interaction terms were significantly different from zero. A partial F-test was used again, with *Model2.(b)* as the full model and *Model2.(b)* without the interaction parameters as the reduced model. The null hypothesis is that all the added interaction parameters are equal to zero. The test statistic was 5.67 and the P-value was $0.0038 < 0.05$ and once again the null hypothesis was rejected. With H_0 true the distribution would be F(2, 290-(2+1)).

Model2.(b) was used to calculate 95 % confidence intervals for the expected log-PM_{10} and PM_{10} for each of the 6 part/coastal combinations when Vehicles = 1000 vehicles per 1000 inhabitants, the results can be seen in table 5.

Table 5: The 95 % confidence intervals for the expected log-PM_{10} and PM_{10} when Vehicles = 1000 per 1000 inhabitants.

Part	Coastal	log-PM_{10} interval	PM_{10} interval
Svealand	Yes	(1.383, 1.519)	(3.986, 4.570)
Svealand	No	(1.470, 1.742)	(4.350, 5.711)
Götaland	Yes	(1.482, 1.605)	(4.4034, 4.977)
Götaland	No	(1.493, 1.667)	(4.451, 5.299)
Norrland	Yes	(1.166, 1.398)	(3.208, 4.045)
Norrland	No	(1.523, 1.805)	(4.585, 6.091)

c

To minimize the amount of variables used the different Part/Coastal combinations were looked at to see if any of them could replace all of the combinations used in the model. First the reference category was set to Coastal = "Yes". Then different combinations of Part/Coastal were tested against the full model, *Model.2(b)* with partial F-tests to see if all the combinations could be replaced by one/some of them. After testing a lot of different combinations we found that the best one to use, with a P-value over 0.05 but as close to 0.05 as possible, was Part= "Gotaland", Coastal="No", Part="Norrland", Coastal="Yes" and Part="Svealand", Coastal = "Yes". So either a municipality is in Götaland/inland-Svealand/Götaland, in inland-Svealand or inland-Norrland. With this combination the P-value of the test was 0.4179 which means there was no significant difference between the full and reduced model.

The new variable NewParts was factorized into "GotalandYes", "SvealandNo" and "NorrlandNo" to represent the combinations mentioned above and a new model was created, *Model.2(c)*. The model is

$$\ln(\text{PM}_{10}) = \beta_0 + \beta_1 \ln(\text{Vehicles}) + \beta_{\text{SvealandNo}} \cdot \text{NewParts} + \beta_{\text{NorrlandNo}} \cdot \text{NewParts} + \epsilon.$$

Table 6: The β-estimates and confidence intervals for *Model.2(c)*.

β-coefficient	Estimate	Lower interval	Upper interval
β_0	-8.5012	-9.444	-7.559
β_1	1.458	1.320	1.597
$\beta_{\text{SvealandNo}}$	-0.1213	-0.2020	-0.04053
$\beta_{\text{NorrlandNo}}$	-0.2892	-0.4181	-0.1603

d

To include other variables in our model we looked at the other numerical variables in our data. All of them were plotted against log-PM_{10} both logarithmic and non logarithmic to see if we should take the logarithm of any of the variables. The plots can be seen in figure 7-12.

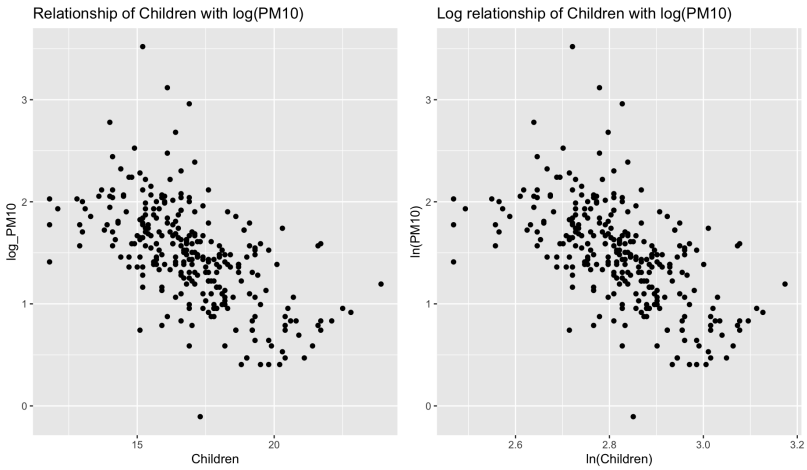


Figure 7: $\ln(PM_{10})$ against Children and ln(Children).

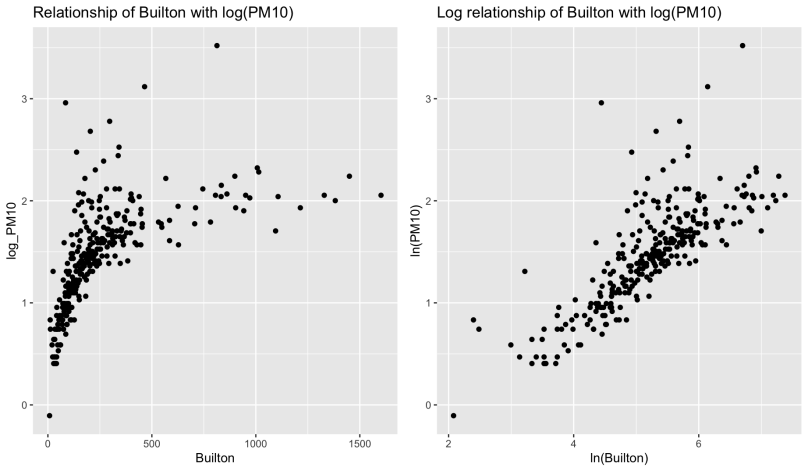


Figure 8: $\ln(PM_{10})$ against Builton and ln(Builton).

Figure 9: $\ln(PM_{10})$ against GRP and ln(GRP).

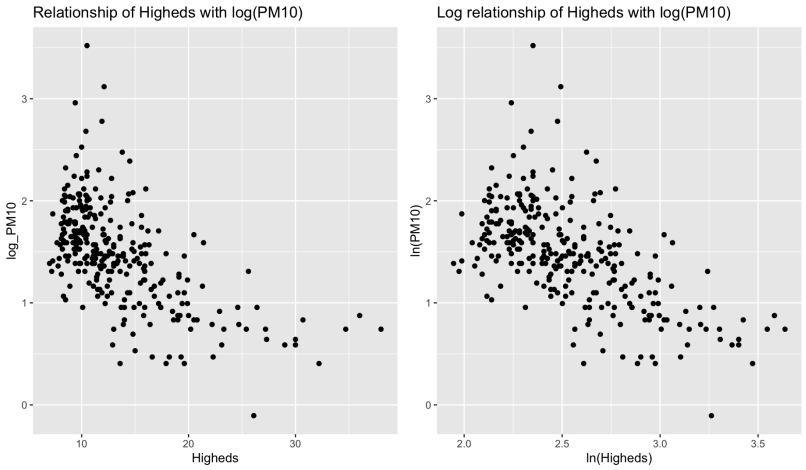


Figure 10: $\ln(PM_{10})$ against Higheds and ln(Higheds).

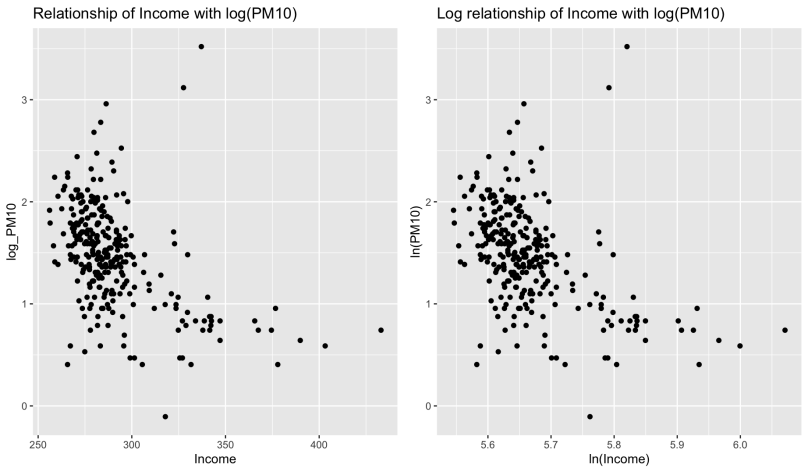


Figure 11: $\ln(PM_{10})$ against Income and ln(Income).

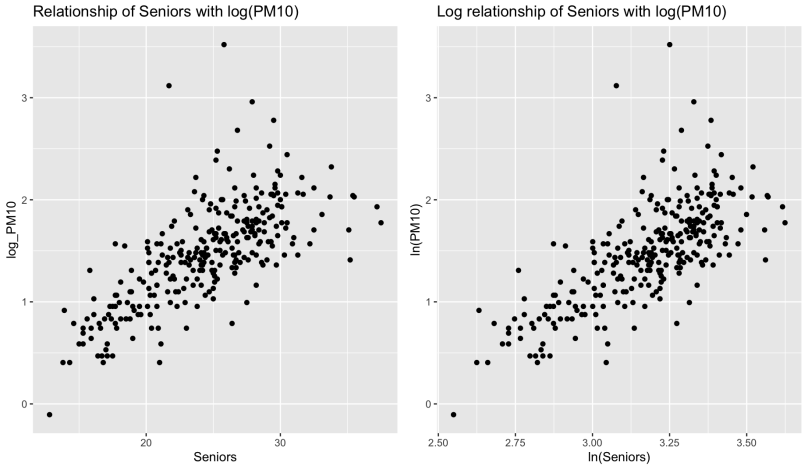


Figure 12: $\ln(PM_{10})$ against Seniors and ln(Seniors).

From those plots we chose the logarithm of the variables Builton, GRP, Income and Higheds. It makes sense to take the logarithm of the economic variables and that seemed to be accurate from the plots as well.

To see which of the numeric variables that are the most correlated with $\ln(PM_{10})$ t-tests were done for Model.1(b) but with each of the numeric variables added one at a time. It was done to see which of the variables were most significant and should be included in the model. From doing the t-tests for all models with a different parameter added the two parameters with the smallest P-values were Higheds and Builton. With P-values of 4.57e-05 and 0.000567 respectively. For the model

$$\ln(PM_{10}) = \beta_0 + \beta_1 \ln(\text{Vehicles}) + \beta_2 \ln(\text{Higheds}) + \beta_3 \ln(\text{Builton}) + \epsilon$$

the β coefficients, their standard errors, confidence intervals and the P-values for their t-tests are presented in table 7 and 8.

Table 7: The coefficients, standard error, t-value and P-value for the model presented above.

Coefficient	Estimate	Std. Error	t value	P-value
β_0	-4.1800	0.8847	-4.726	3.61e-06
β_1	0.8073	0.16810	4.802	2.53e-06
β_2	-0.21461	0.07357	-2.917	0.00381
β_3	0.12260	0.06392	1.918	0.05612

Table 8: The β-estimates confidence intervals for the model above.

β-coefficient	Lower interval	Upper interval
β_0	-5.922	-2.439
β_1	0.4764	1.138
β_2	-0.3594	-0.06981
β_3	-0.003221	0.2484

For the three variables the VIF values were calculated and are presented in table 9. All the variables were also plotted against each other as well their correlations in figure 13.

Table 9: The VIF values for the variables in the model.

Variable	VIF-value
Vehicles	8.5462
Higheds	2.005
Builton	10.30

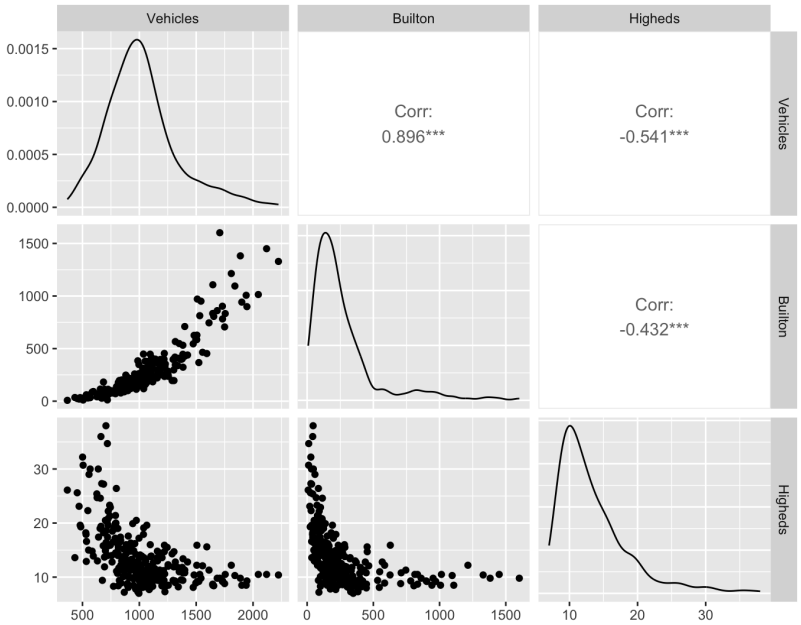


Figure 13: The correlation between the three x-variables.

All three variables have relatively low VIF-values but the highest one is Builton. Looking at the correlation plot and values it is shown that the absolute correlation value is the highest for Builton with the two other variables. From that information it looks like Builton could be a problematic variable to use together with Vehicles and Higheds. We chose to exclude the Builton-variable and our new model, *Model.2(d)* is

$$\ln(\mathrm{PM}_{10}) = \beta_0 + \beta_1 \ln(\mathrm{Vehicles}) + \beta_2 \ln(\mathrm{Higheds}) + \epsilon.$$

In table 10 the parameter estimates, standard errors and P-values are presented. In table 11 the confidence intervals are presented.

Table 10: The coefficients, standard error, t-value and P-value for *Model.2(d)*.

Coefficient	Estimate	Std. Error	P-value
β_0	-5.383	0.6271	5.86e-16
β_1	1.097	0.07372	$<$ 2e-16
β_2	-0.2757	0.06661	4.57e-05

Table 11: The β-estimates confidence intervals for *Model.2(d)*.

β-coefficient	Lower interval	Upper interval
β_0	-6.618	-4.149
β_1	0.9522	1.242
β_2	-0.4067	-0.1445

The new VIF values are 1.628 for Vehicles and 1.628 for Higheds. They are much lower now compared to when the Builton variable was also included in the model. $\sqrt{\text{VIF}_j}$ indicates how many times larger the standard error is because of the dependence with the other x-variables. The VIF value for Vehicles was 8.5462 with Builton in the model, $\sqrt{8.5462} = 2.9233$ which means the standard error should be 2.9233 times higher for Vehicles with Builton in the model. The standard error for log-Vehicles was 0.168 with Builton in the model and 0.074 without the Builton variable which is a 2.270 increase with Builton is the model. This value is pretty close to $\sqrt{\text{VIF}_{\text{Vehicles}}}$ which means most of the dependency was with the Builton variable and not the Higheds variable.

e

Now we create another model that includes all the continuous variables except Builton that was excluded previously, the variable NewParts is also included in this model. Our new model is

$$\begin{aligned}\text{PM}_{10} =&\beta_0 + \beta_1 \ln(\text{Vehicles}) + \beta_2 \ln(\text{Higheds}) + \beta_3 \text{Children} + \beta_4 \text{Seniors} + \\ &\beta_5 \ln(\text{Income}) + \beta_6 \ln(\text{GRP}) + \beta_7 \text{NewParts} + \epsilon.\end{aligned}$$

The new model was studied by looking at the VIF values which are now GVIF values since we have included categorical variables. The GVIF values are presented in table 12. The x-variables were also plotted against each other, see figure 14

Table 12: The GVIF values for the variables in the model above.

Variable	GVIF-value
Vehicles	2.957
Higheds	3.008
Children	5.211
Seniors	7.030
Income	2.652
GRP	1.447
NewParts	1.660

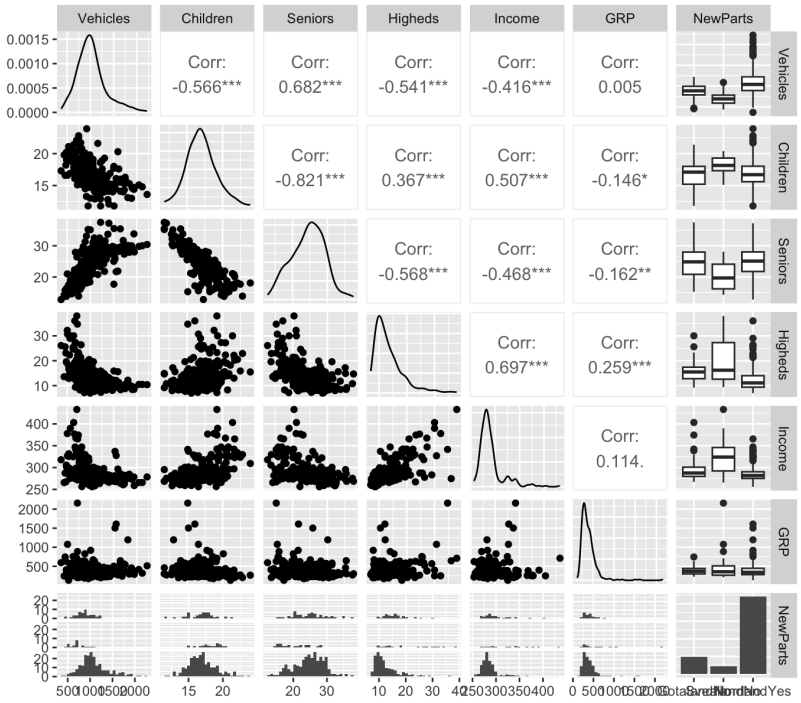


Figure 14: The correlation between the variables in the model presented above.

The Seniors variable has the highest GVIF value. In figure 14 the correlation between all the variables is plotted and from the figure it is clear that the Seniors variable has high correlation with almost all the other variables, especially with Children where $\rho = -0.821$. From both the plot and the GIF values it seems like Seniors could be a problematic variable to include in the model since it is highly correlated to the other variables. It should definitely not be used in a model with the Children variable included. The Seniors variable is excluded from the model and the new model, *Model.2(e)* is

$$\begin{aligned} \text{PM}_{10} =& \beta_0 + \beta_1 \ln(\text{Vehicles}) + \beta_2 \ln(\text{Higheds}) + \beta_3 \text{Children} \\ &+ \beta_4 \ln(\text{Income}) + \beta_4 \ln(\text{GRP}) + \beta_4 \text{NewParts} + \epsilon. \end{aligned}$$

The new GVIF values are presented in table 13 and they are all lower now that Seniors was excluded from the model.

Table 13: The GVIF values for the variables in *Model.2(e).*

Variable	GVIF-value
Vehicles	2.197
Higheds	2.795
Children	1.999
Income	2.242
GRP	1.141
NewParts	1.281

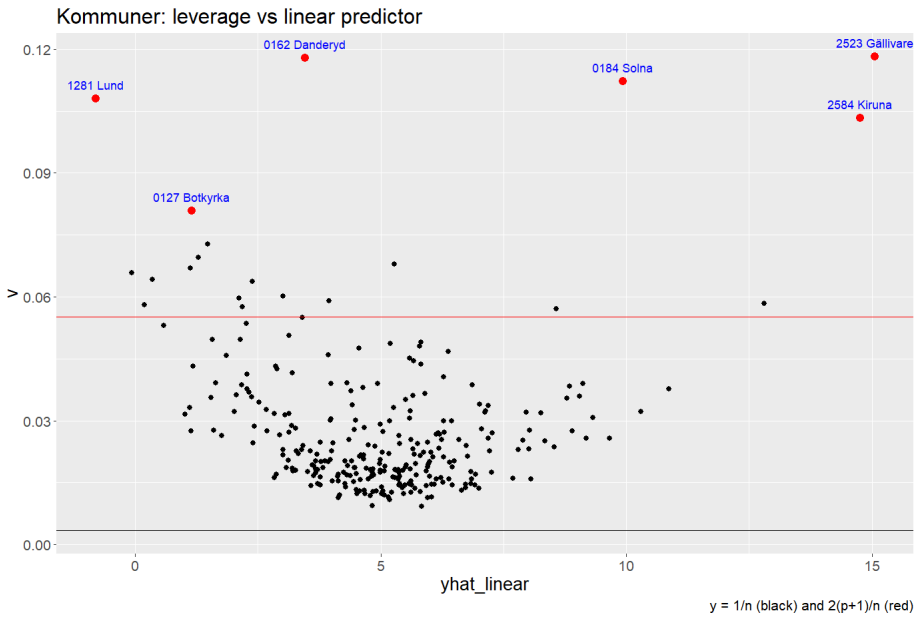
Part 3: Model validation and selection

a

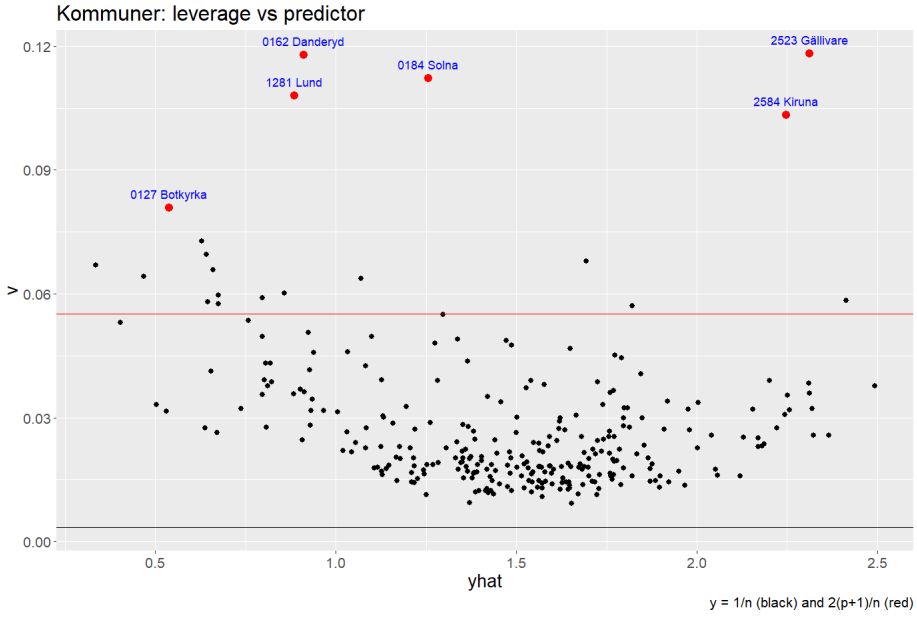
To make the model more stable and reliable, we need to preprocess the data. The leverage can help identify data points that may have a disproportionate influence on the model estimates in linear regression. Here, we compare the leverage values with $\frac{1}{n}$ and $\frac{2(p+1)}{n}$ to determine whether a data point is likely to affect the stability and reliability of the regression line. It also facilitates the subsequent cleaning of potential outlier data points.

In figures 15a and 15b, we identified the six municipalities with the highest leverage and highlighted them.

The linear model that we used to obtain the linear prediction results is *Model.2(e).*



(a) Linear model.



(b) Model 2(e).

Figure 15: The plot of leverage from *Model 2(e)* against the linear predictor and non-linear predictor.

Plotting pairs of x-variables against each other, separately for each factor category, is shown in figure 16.

From figure 16, we can observe that municipalities with higher leverage values are distributed in the NorrlandYes category, except for Lund. In other words, almost every municipality with higher leverage are located in Norrland. Hence, it is evident that this particular x-variable significantly contributes to higher leverage values.

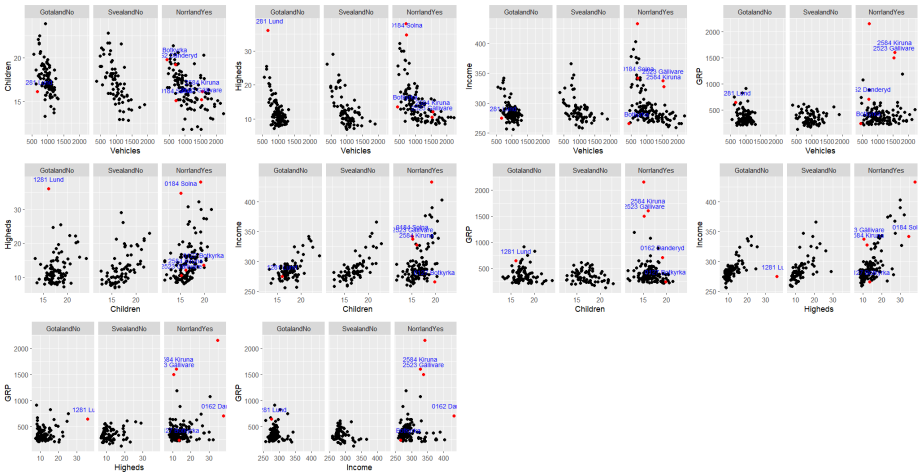


Figure 16: The plots depicting the relationships between x-variables, only the plots between different x-variables were created, and the data points with high leverage were highlighted in red.

If we focus on the distribution of outliers and examine the relationship between a specific x-variable and other x-variables, it becomes evident that data points with higher leverage values are noticeably prominent outliers in the plots of GRP against other x-variables. This indicates that the parameter GRP is likely causing the higher leverage.

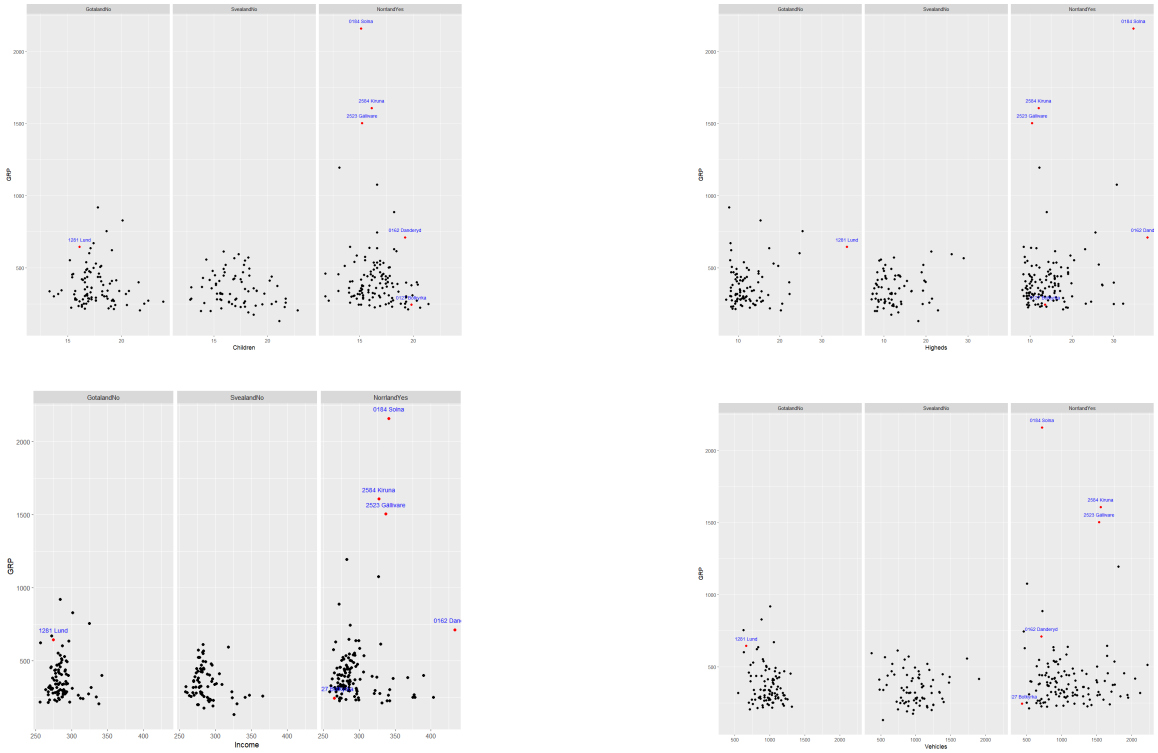


Figure 17: The plots of GRP paired with other x-variables, the data points with high leverage were highlighted in red.

b

Cook's distance is also a statistical measure used to assess the potential impact of data points on a model. By analyzing Cook's distance, we can estimate the overall transformation of the model after removing a specific data point. If Cook's distance exceeds a critical value, we consider that the data point may have a significant influence on the model and further investigation is warranted.

Here, we use two parameters to draw suitable horizontal lines as visual references, namely $F_{0.5,p+1,n-(p+1)}$ and $4/n$, as shown in Figure 18. It can be observed from the figure that $F_{0.5,p+1,n-(p+1)}$ is a conservative approach and represents the value at the median of the F-distribution with $p+1$ and $n-(p+1)$ degrees of freedom, where p is the number of predictors. None of the D_i values in the entire dataset exceed F. On the other hand, data points with D_i values below $4/n$ are considered to have minimal impact on the model.

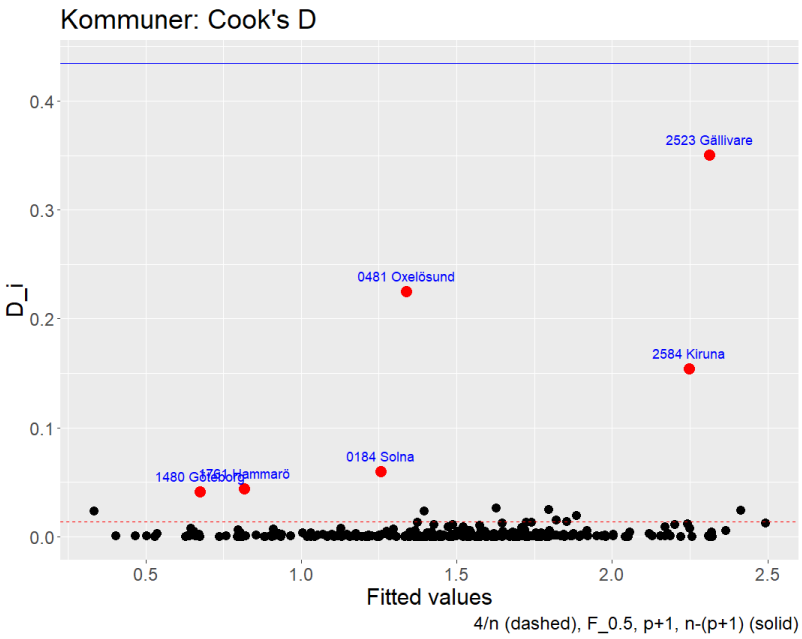


Figure 18: Cook's distance against the linear predictor.

We also identify the six municipalities with the highest Cook's distances and highlight them in red on the plot for clearer visibility. These municipalities are then compared with the six that exhibit the highest leverage. The comparative results are detailed in Table 14.

Table 14: Comparison of Municipalities with the Highest Cook's Distance and Leverage

Rank	Top Cook's Distance	Top Leverage
1	2584 Kiruna	2523 Gällivare
2	2523 Gällivare	0162 Danderyd
3	1761 Hammarö	0184 Solna
4	1480 Göteborg	1281 Lund
5	0481 Oxelösund	2584 Kiruna
6	0184 Solna	0127 Botkyrka

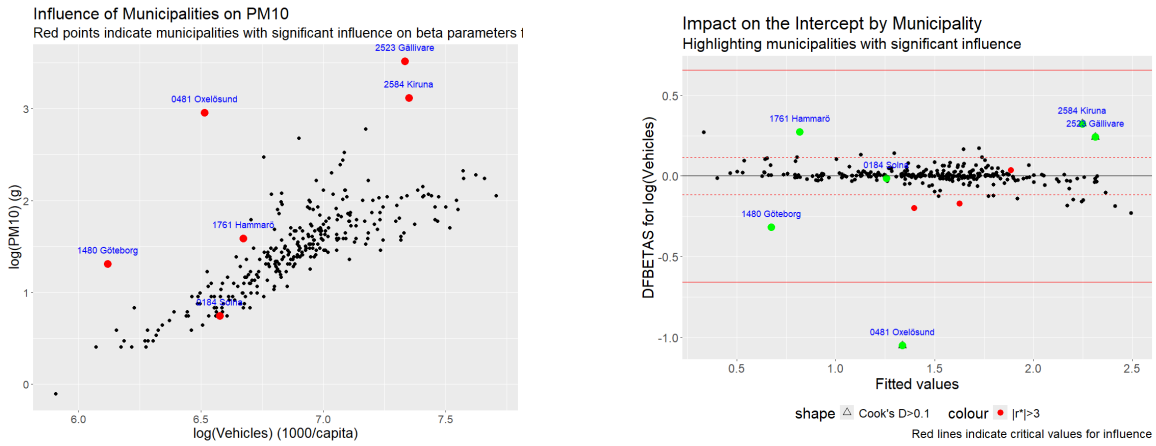
Upon comparison, it is evident that the municipalities of Gällivare, Kiruna, and Solna appear in both lists, indicating their significant impact on the model across both statistical measures. This dual presence underscores the importance of these municipalities in the analysis and suggests that they may be critical points for further detailed investigation.

We obtained 6 municipalities with Cook's distances and calculated their DFBETA values with different x-variables, as shown in Table 15

We contend that the concurrent use of DFBETAS plots and graphs of log-PM_{10} against relevant variables offers a better methodology for demonstrating and substantiating the impact of data points with high Cook's distances on the β parameters. This approach effectively highlights the influence of these points, providing a clear and robust verification of their impact on the model.

Table 15: Top Cook's Distance Municipalities with DFBETAS Values

Kommun	D	Intercept	log(Vehicles)	log(Higheds)	Children	log(Income)	log(GRP)
Gällivare	0.494	-1.40	-0.078	-0.968	-0.204	1.33	1.34
Kiruna	0.217	-0.829	0.138	-0.409	0.094	0.647	0.985
Oxelösund	0.201	0.0495	-0.995	-1.04	-0.326	0.401	0.478
Solna	0.08	0.288	-0.075	-0.172	0.166	-0.164	-0.571
Göteborg	0.053	0.336	-0.373	0.155	-0.184	-0.218	0.159
Hammarö	0.046	-0.086	0.244	0.268	0.366	-0.069	-0.033

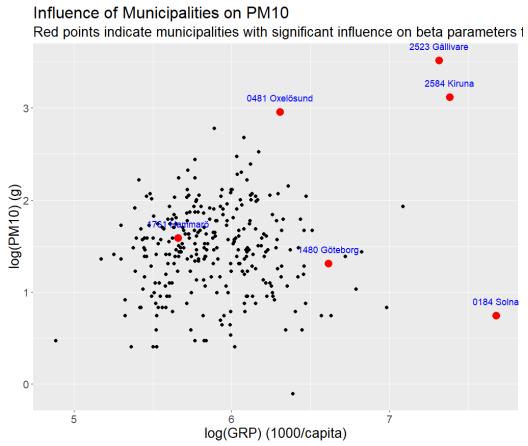


(a) $\log(PM_{10})$ against log(Vehicles)

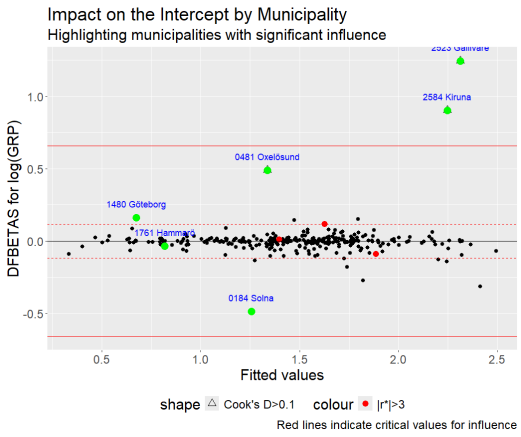
(b) DFBETAS plots of log(Vehicles)

Figure 19: Figure of $\log(PM_{10})$ against log(Vehicles) and DFBETAS plot.

From the analyses presented in the two preceding plots, it is apparent that Solna does not markedly deviate in the plot of $\log(PM_{10})$ against log(Vehicles), and it exhibits relatively small absolute values of $|\text{DFBETAS}_{j(i)}|$ in the DFBETAS plot. Conversely, other data points with higher Cook's distances, notably Oxelösund, show larger absolute values of $|\text{DFBETAS}_{j(i)}|$, indicative of their significant influence. Furthermore, Oxelösund is distinctly problematic in the plot of $\log(PM_{10})$ against log(Vehicles), suggesting a pronounced impact on the model's behavior.



(a) $\log(PM_{10})$ against log(GRP)

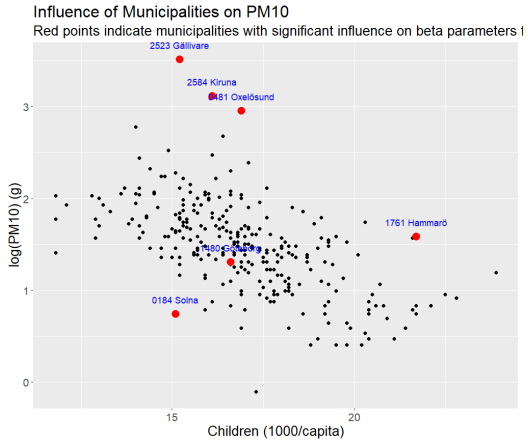


(b) DFBETAS plots of log(GRP)

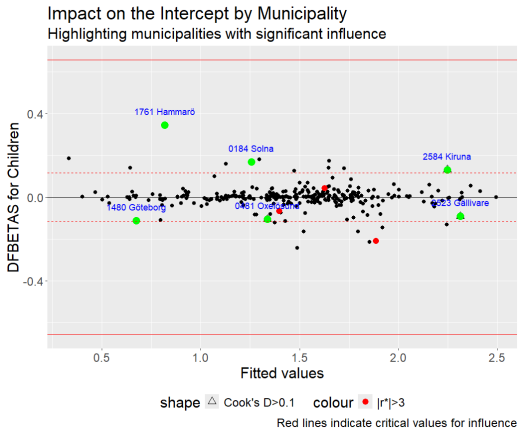
Figure 20: Figure of $\log(PM_{10})$ against log(GRP) and DFBETAS plot.

From figure 20, it is evident that Gällivare, Kiruna, Oxelösund, and Solna emerge as prominent outliers when the x-variable is log(GRP). This observation is corroborated by the DFBETAS plot, where these four municipalities display significantly large absolute values of $|\text{DFBETAS}_{j(i)}|$. This leads to the conclusion that log(GRP) is the x-variable most impacted by municipalities exhibiting the highest Cook's distances.

Based on figure 21, it can be observed that municipalities with the highest Cook's distances exert a minimal impact on the β parameter for the variable "Children". The plot of $\log(PM_{10})$ against Children does not exhibit prominent outliers. Additionally, the DFBETAS plot reveals no data points with significant absolute values of $\text{DFBETAS}_{j(i)}$ for this variable.

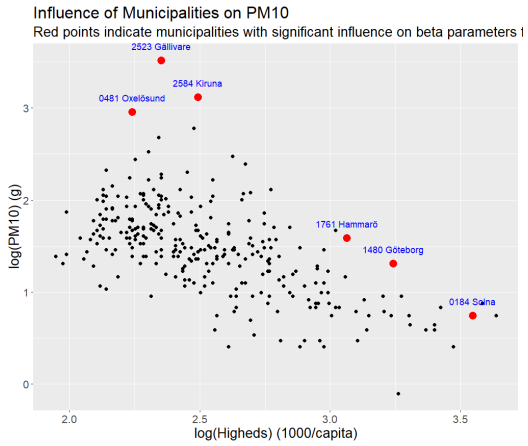


(a) $\log(PM_{10})$ against Children

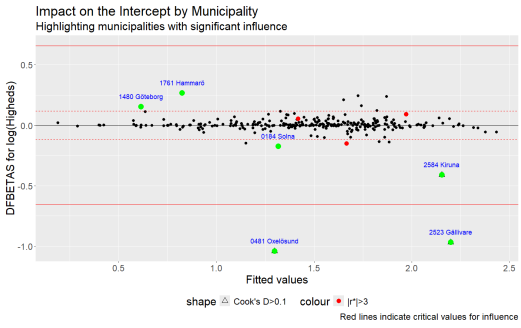


(b) DFBETAS plots of Children

Figure 21: Figure of $\log(PM_{10})$ against Children and DFBETAS plot.

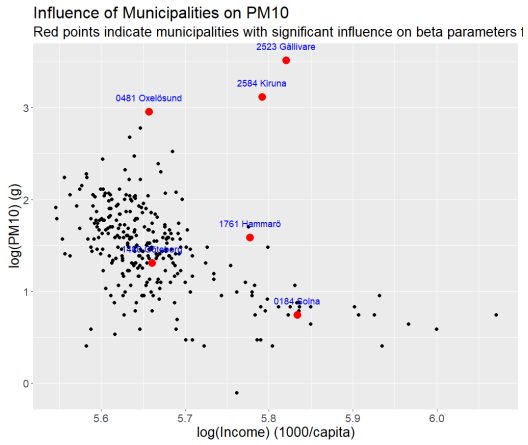


(a) $\log(PM_{10})$ against log(Higheds)

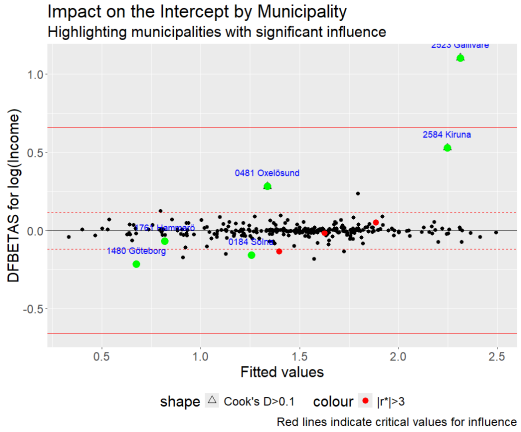


(b) DFBETAS plots of log(Higheds)

Figure 22: Figure of $\log(PM_{10})$ against log(Higheds) and DFBETAS plot.



(a) $\log(PM_{10})$ against log(Income)



(b) DFBETAS plots of log(Income)

Figure 23: Figure of $\log(PM_{10})$ against log(Income) and DFBETAS plot.

As depicted in Figure 22, there are two prominent outliers in the plot of log(Higheds), and Figure 23 reveals one significant outlier in log(Income). Although the β parameters for these variables are influenced by the municipalities with the highest Cook's distances, the impact is not the most substantial. Based on the analysis, it is evident that log(GRP) is the x-variable most significantly affected by municipalities with the highest Cook's distances.

c

The studentized residual, denoted as r_i^*, is a standardized residual utilized extensively in linear regression to pinpoint outliers and influential data points. The calculation for the studentized residual is given by:

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}}$$

In practice, studentized residuals are instrumental in detecting potential outliers. Commonly employed thresholds are ± 2 and ± 3 for identifying outliers. Observations with studentized residuals exceeding ± 3 are considered highly exceptional and are strong candidates for outliers. These extreme values can significantly impact the regression model and may lead to distortions in the estimation of regression coefficients.

From Figure 24, the municipalities with the highest Cook's distances are highlighted in red. A total of three municipalities exhibit both high Cook's distances and high studentized residuals, with $|r_i^*| > 3$. These municipalities are Oxelösund, Gällivare, and Kiruna. The municipalities of Karlshamn, Monsteras, and Kalix do not have the largest Cook's distance, but their residuals r_i^* are greater than 3, they were highlighted in blue.

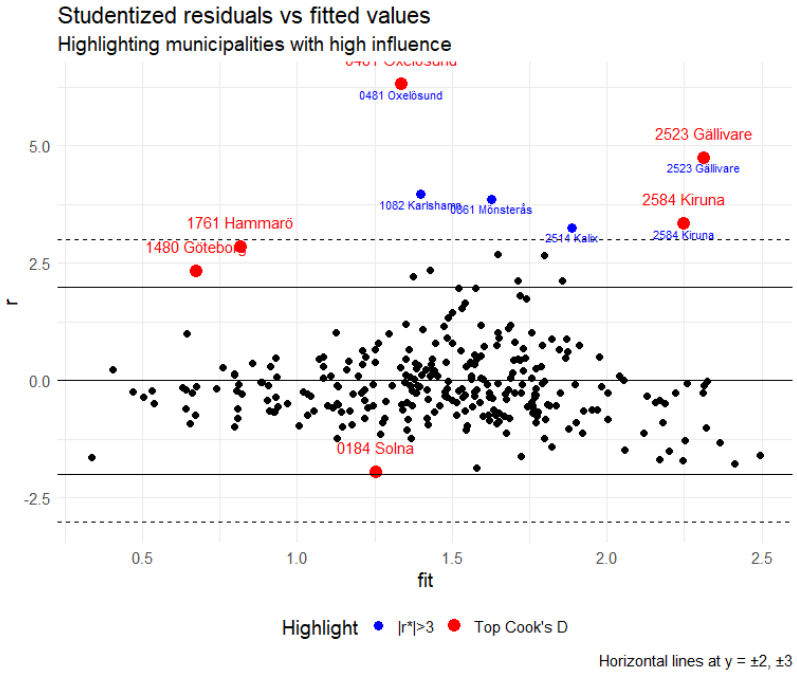


Figure 24: Studentized residuals against the linear predictor.

Plot of $\sqrt{|r_i^*|}$ against the linear predictor is shown in figure 25. Based on the analysis of the plot, it appears that the variance remains constant across the range of predicted values.

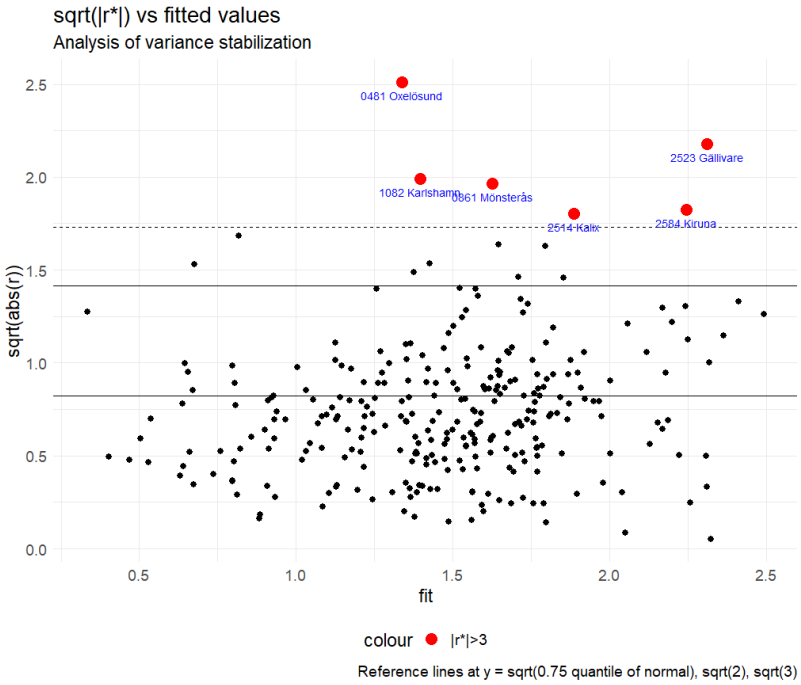
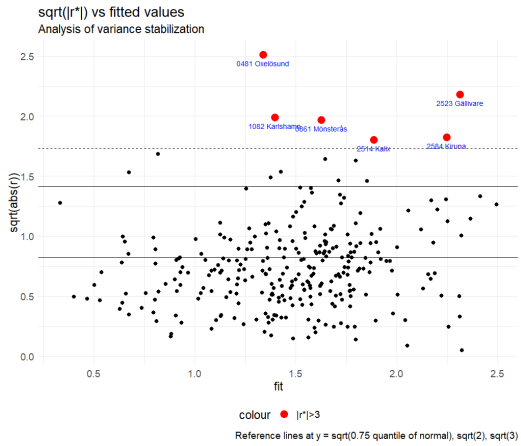


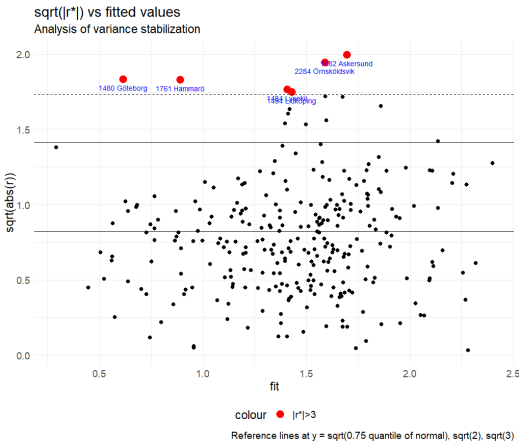
Figure 25: $\sqrt{|r_i^*|}$ against the linear predictor.

d

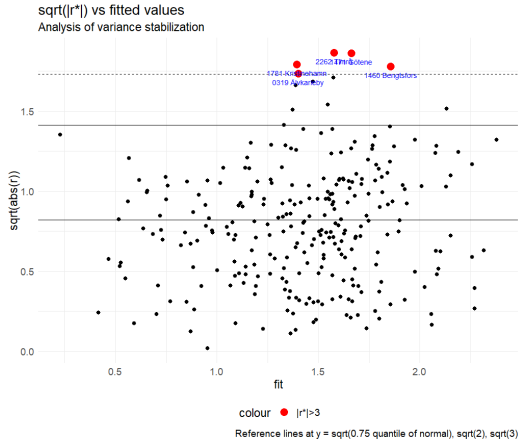
The municipalities exhibiting high residuals are depicted in figure 26, with subfigures (a) $\sim$ (d) representing the respective municipalities at each of the four sequential steps of analysis. Comprehensive descriptions of these high-residual municipalities, as identified in each step, are detailed in Table 16 $\sim$ 19 .



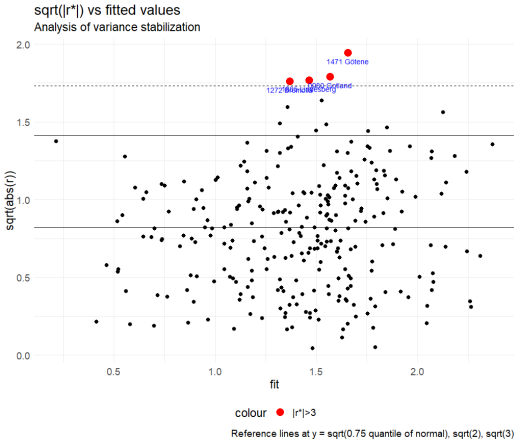
(a) High-residual municipalities that $r_i^* > 3$ in step 1.



(b) High-residual municipalities that $r_i^* > 3$ in step 2.



(c) High-residual municipalities that $r_i^* > 3$ in step 3.



(d) High-residual municipalities that $r_i^* > 3$ in step 4.

Figure 26: High-residual municipalities across four steps

In the final plot of the new data, as shown in figure 27, there is still a municipality, Götene, with a high residual. However, we did not remove it because we could not determine its pollution source.

Table 16: Municipalities and Associated Companies excluded in step 1

Municipality Code	Company Name	Company Type
0481 Oxelösund	SSAB	Steel company
	SMA Mineral AB	Lime company
1082 Karlshamn	Södra Cell	Paper mills
0861 Mönsterås	Södra Cell	Paper mills
2523 Gällivare	Luossavaara-Kiirunavaara AB (LKAB)	State owned mining company
	Boliden Mineral AB	Mining company
2514 Kalix	Billerud AB	Paper mills
2584 Kiruna	Luossavaara-Kiirunavaara AB (LKAB)	State owned mining company

Table 17: Municipalities and Associated Companies excluded in step 2

Municipality Code	Company Name	Company Type
1480 Göteborg	Preem AB	Petroleum and bio-fuel company with oil refinery
	St1 Refinery AB	Oil refinery
	Renova AB	Garbage/waste/recycling company
	The Swedish Air force wing F-7	
1761 Hammarö	Stora Enso	Petroleum and bio-fuel company
1484 Lysekil	Preem AB	Paper mill
1494 Lidköping	The Swedish Air force wing F-7	
1882 Askersund	Ahlstrom	Paper mill
2284 Örnsköldsvik	Metsä Board Sverige AB	Paper mill

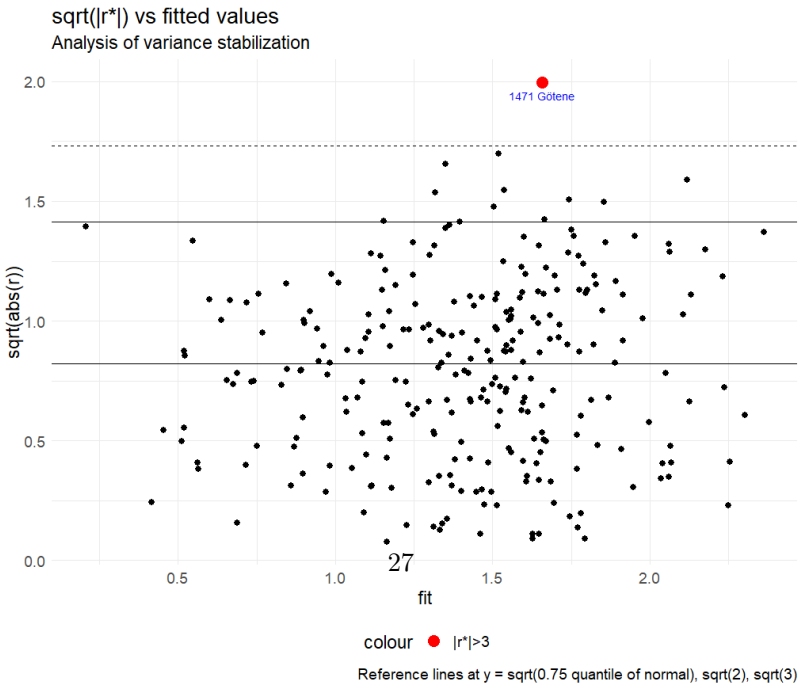


Figure 27: Final high-residual municipality that $r_i^* > 3$.

Table 18: Municipalities and Associated Companies in Step 3

Municipality Code	Company Name	Company Type
0319 Älvkarleby	Stora Enso	Paper mills
1460 Bengtsfors	Ahlstrom	Paper mills
1781 Kristinehamn	Nordic Paper	Paper mills
2262 Timrå	SCA	Paper mills

Table 19: Municipalities and Associated Companies in Step 4

Municipality Code	Company Name	Company Type
0980 Gotland	Cementa AB / Heidelberg Materials	Building materials company
1272 Bromölla	Stora Enso	Paper mills
1885 Lindesberg	Billerud AB	Paper mills
1764 Grums	Billerud AB	Paper mills

Tables 20 and 21 display the model statistical metrics after removing outliers and the parameters for *Model (2e)*, respectively. We now proceed to analyze these statistical metrics to assess the influence of outliers on the coefficients of the model. Our primary focus includes:

- Standard Error: This metric reflects the precision of the estimates. A smaller standard error suggests reduced uncertainty and enhanced stability of the model.
- T-value and p-value: These statistics are used to test the significance of the variables within the model. Higher absolute t-values along with smaller corresponding p-values indicate a stronger statistical significance of the variable.
- Confidence Interval (95% CI): This indicates the accuracy of the parameter estimates. A narrower interval suggests a more precise estimate.

By comparing the data from the two tables, it is evident that the standard errors for all coefficients have decreased after removing outliers. This reduction in standard errors signifies enhanced model stability and diminished uncertainty. For instance, the standard error for the coefficient of log(Vehicles) decreased from 0.09294 to 0.05424.

Upon comparing the two models, notable changes in certain coefficients' t-values and p-values are evident after outlier removal, affecting their significance. For instance, in *Model 2e*, the t-value for the intercept increased from -3.909 to -2.263, while its corresponding p-value decreased. This suggests that the intercept is no longer statistically significant in the refined model. Conversely, for log(Vehicles), the t-value rose from 12.044 to 22.194, indicating strengthened significance in the refined model. Similar shifts are observed for other coefficients. Similar results can be observed regarding the analysis of confidence intervals.

In the refined model, variables like log(Vehicles), log(Higheds), and Children exhibit heightened significance, as indicated by higher t-values and lower p-values. This implies their increased explanatory power following outlier removal. Conversely, the significance of the intercept and log(Income) has diminished, evidenced by lower t-values and higher p-values. This suggests that their impact in the refined model might not be as substantial.

The significance of log(GRP) has decreased significantly, but the standard deviation has also decreased. This confirms our previous observation that outliers have a substantial impact on log(GRP) and, consequently, on the overall model. This is why log(GRP) initially exhibited high significance. However, after removing outliers, the significance of log(GRP) has decreased.

Table 20: Regression Statistics for the Excluded Municipalities Model

Coefficient	Estimate	Std. Error	t value	Pr(> |t|)	95% CI
Intercept	-2.363069	1.044365	-2.263	0.024474 *	(-4.4195, -0.3067)
log(Vehicles)	1.203881	0.054245	22.194	$< 2 \times 10^{-16}$ ***	(1.0971, 1.3107)
log(Higheds)	-0.108466	0.047523	-2.282	0.023269 *	(-0.2020, -0.0149)
Children	-0.018279	0.006201	-2.948	0.003490 **	(-0.0305, -0.0061)
log(Income)	-0.684905	0.183974	-3.723	0.000241 ***	(-1.0472, -0.3226)
log(GRP)	0.003515	0.029017	0.121	0.903688	(-0.0536, 0.0607)
NewParts Svealand No	-0.123266	0.022360	-5.513	8.45×10^{-8} ***	(-0.1673, -0.0792)
NewParts Norrland No	-0.274490	0.036141	-7.595	5.45×10^{-13} ***	(-0.3456, -0.2033)

Table 21: Regression Statistics for *Model 2(e)*

Coefficient	Estimate	Std. Error	t value	Pr(> |t|)	95% CI
Intercept	-6.69945	1.71366	-3.909	0.000116 ***	(-10.0726, -3.3263)
log(Vehicles)	1.11931	0.09294	12.044	$< 2 \times 10^{-16}$ ***	(0.9364, 1.3022)
log(Higheds)	-0.31051	0.07897	-3.932	0.000106 ***	(-0.4660, -0.1551)
Children	-0.03248	0.01073	-3.027	0.002702 **	(-0.0536, -0.0114)
log(Income)	0.12297	0.30563	0.402	0.687732	(-0.4786, 0.7246)
log(GRP)	0.19889	0.04737	4.198	3.61×10^{-5} ***	(0.1056, 0.2921)
NewParts Svealand No	-0.12624	0.03833	-3.294	0.001116 **	(-0.2017, -0.0508)
NewParts Norrland No	-0.25793	0.06263	-4.118	5.02×10^{-5} ***	(-0.3812, -0.1346)

e

The results of stepwise selection using AIC and BIC as criteria are shown in Table 22a and Table 22b, respectively. These tables include each step and indicate which variables were included or excluded from the model at each step.

Table **??** contains the number of β-parameters, the residual standard deviation, the R2, adjusted R2, AIC, and BIC for the six models.

(a) Stepwise Regression Analysis by AIC

Step	Include/Exclude Variable	AIC
Start	$\log(PM_{10}) \sim \log(Vehicles)$	-887.82
1	+ NewParts	-945.40
2	+ log(Income)	-998.11
3	+ Children	-1002.85
4	+ log(Higheds)	-1006.36

(b) Stepwise Regression Analysis by BIC

Step	Changes in Model	AIC
Start	$\log(PM_{10}) \sim \log(Vehicles)$	-880.63
1	+ NewParts	-931.01
2	+ log(Income)	-980.12
3	+ Children	-981.26

Table 22: Stepwise Regression Analysis by AIC and BIC

Table 23: Statistical Models Comparison

Model	β-Parameters	Residual SD	R^2	R^2_{adj}	AIC	BIC
Null	1	0.4364	0.0000	0.0000	321.4386	328.6355
Model 1(b)	2	0.1925	0.8062	0.8055	-119.5962	-108.8010
Model 2(c)	4	0.1724	0.8457	0.8440	-177.1741	-159.1820
Model 3(d)	8	0.1534	0.8796	0.8764	-236.1455	-203.7597
AIC Model	7	0.1531	0.8796	0.8769	-238.1304	-209.3430
BIC Model	6	0.1544	0.8771	0.8748	-234.6262	-209.4372

The choice of the best model depends on the balance between model complexity (number of parameters) and goodness of fit (R-squared, adjusted R-squared). *Model 3(d)* has the highest R-squared and adjusted R-squared values among all models, indicating it explains the most variability in log-PM_{10} emissions while considering model complexity. However, the AIC and BIC models achieve similar goodness-of-fit metrics with fewer parameters, suggesting potential overfitting in *Model 3(d)*. Therefore, either the AIC or BIC model may be considered the best, depending on the preference for model simplicity or fit. I would choose BIC model as the best model since it has less model complexity but similar performance.

The BIC model is

$$\begin{aligned}\ln(\text{PM}_{10}) =& \beta_0 + \beta_1 \ln(\text{Vehicles}) + \beta_2 \cdot \text{Children} + \beta_3 \cdot \ln(\text{Income}) + \\ & \beta_{\text{SvealandNo}} \cdot \text{NewParts} + \beta_{\text{NorrlandNo}} \cdot \text{NewParts} + \epsilon.\end{aligned}$$

Table 24: Regression Coefficients for BIC Model

Coefficient	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-1.652358	0.944979	-1.749	0.0815
log(Vehicles)	1.269869	0.046130	27.528	$< 2 \times 10^{-16}$ ***
NewPartsSvealandNo	-0.118400	0.022292	-5.311	2.31×10^{-7} ***
NewPartsNorrlandNo	-0.293329	0.035304	-8.309	5.13×10^{-15} ***
Children	-0.015421	0.005969	-2.583	0.0103 *
log(Income)	-0.943900	0.145619	-6.482	4.42×10^{-10} ***

Model.1(b) indicates that 80.62% of the variability in log-PM_{10} emissions can be explained by

the number of vehicles. However, *Model.1(b)* onwards, there is a significant improvement, with R-squared values increasing steadily. The best model, the BIC model, explains approximately 87.71% of the variability in log-PM_{10} emissions.

Conclusion

The variables included in the models, such as log(Vehicles), log(Income), Children, etc. seem reasonable as potential predictors of log-PM_{10} emissions, as they likely have known relationships with air pollution. The signs of the β-parameters also make sense intuitively. For example, a positive β-parameter for log(Vehicles) suggests that an increase in the number of vehicles is associated with higher log-PM_{10} emissions. Furthermore, the coefficient before log(Vehicles) is larger than the other coefficients, indicating that the number of vehicles has the most significant impact on PM_{10} concentration.

Individuals with higher education levels are more likely to reside in areas with better environmental conditions and lower pollution levels. They may also exhibit a greater concern for environmental conservation and sustainable development. Therefore, the coefficient of log(Higheds) is negative, indicating that areas with higher education levels tend to have lower pollution levels. However, due to the insignificance of the relationship between Higheds and PM_{10} pollution levels compared to other variables, it was omitted from the model for the sake of simplicity.

The coefficient of -0.01542 suggests that an increase in the proportion of children aged 0 to 14 is associated with a decrease in PM_{10} pollution. This may indicate that areas with a higher proportion of children place greater emphasis on environmental protection and implement more pollution control measures.

As income rises, PM_{10} concentration tends to decrease, since the coefficient of -0.943900. This may be because wealthier individuals have the means to adopt eco-friendly practices, like using cleaner energy sources or opting for more sustainable transportation methods. Moreover, those with higher incomes often reside in areas with better environmental conditions, which lowers their exposure to pollution.

In comparison to Götaland, Svealand and Norrland regions experience reduced industrial activity, resulting in lower levels of PM_{10} pollution. Moreover, inland cities generally have lower population and industrial density compared to coastal cities, contributing to decreased pollution levels. When contrasting Svealand with Norrland, the latter typically exhibits lower pollution levels possibly due to its northern location and lower population density.

Use of AI

We used ChatGPT to proofread and polish the report, checking for spelling and grammar errors. ChatGPT was also utilized in the generation process of LaTeX tables, reducing the time required for repetitive tasks.

Author contributions

Both group members were involved in various tasks, including deriving, analyzing, programming, and writing. Elise Olsson took on responsibility for Parts 1 and 2, while Yifan Zhang was responsible for Part 3.