Lund University
Faculty of Engineering MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION
Name: ZHANG Yifei & XU Jialu
Email: yi4840zh-s@lu.se & ji6606xu-s@lu.se
Date: April 24, 2024

Project 1
LINEAR REGRESSION

1 Simple linear regression

1.1 1(a)

1.1.1 Why logarithm:

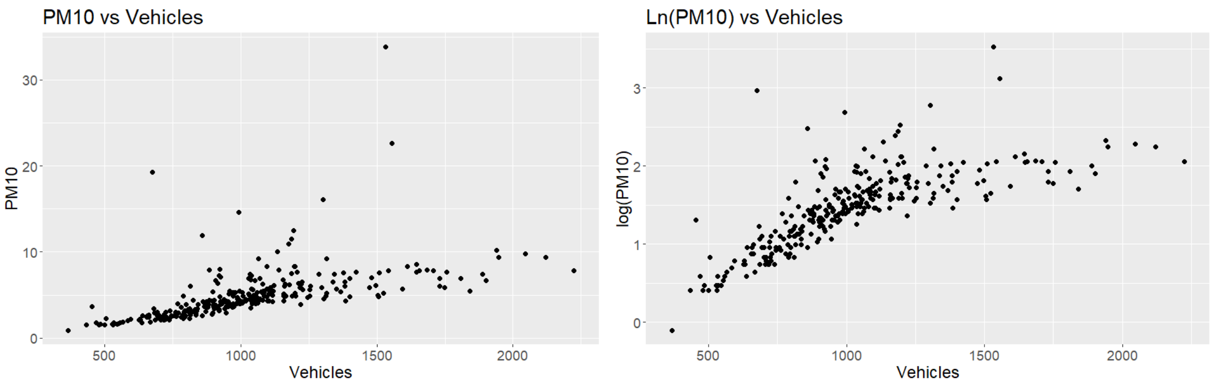


Figure 1: PM10 vs ln(PM10)

Firstly, the scatter plots of PM10 with respect to x and ln(PM10) with respect to x are plotted respectively, from which we can visualize an exponential trend of PM10 with respect to x.

Then we can analyse residual plots for each model:

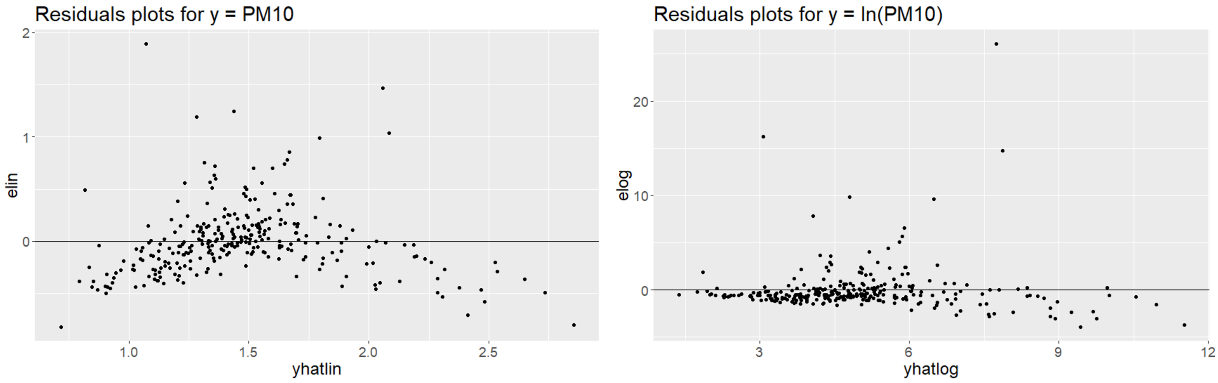


Figure 2: Residual plots for y=PM10(left) and y=ln(PM10)(right)

From the figure, it can be seen that the residuals of the log-transformed model are more consistent with a normal distribution.

This conclusion can also be drawn from the Q-Q plot: as shown below, the Q-Q plot of the residuals of the log-transformed model fits more closely to a straight line.

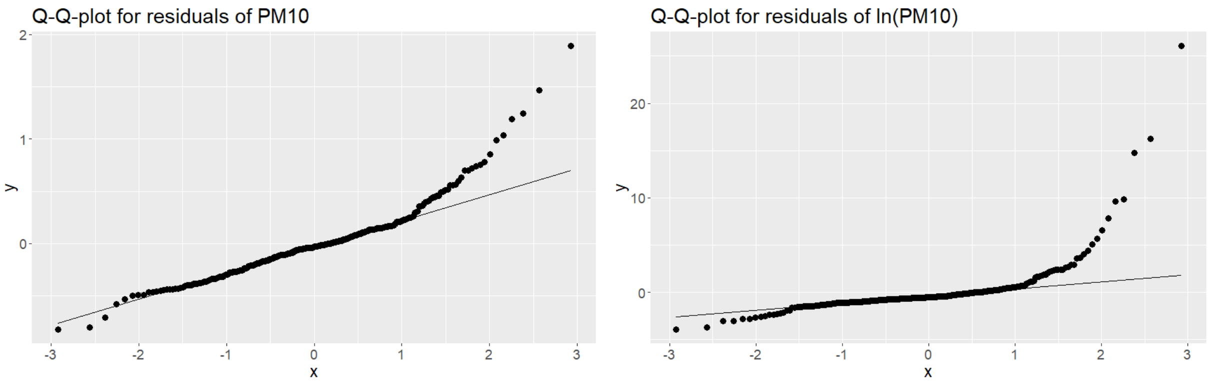


Figure 3: Q-Q plots for y=PM10(left) and y=ln(PM10)(right)

1.1.2 Choose x:

For the x-variable selection problem, similarly, scatter plots of ln(PM10) with respect to x = Vehicles and x = ln(Vehicles), respectively, can be plotted, and it can be observed that the latter has a better degree of linearity.

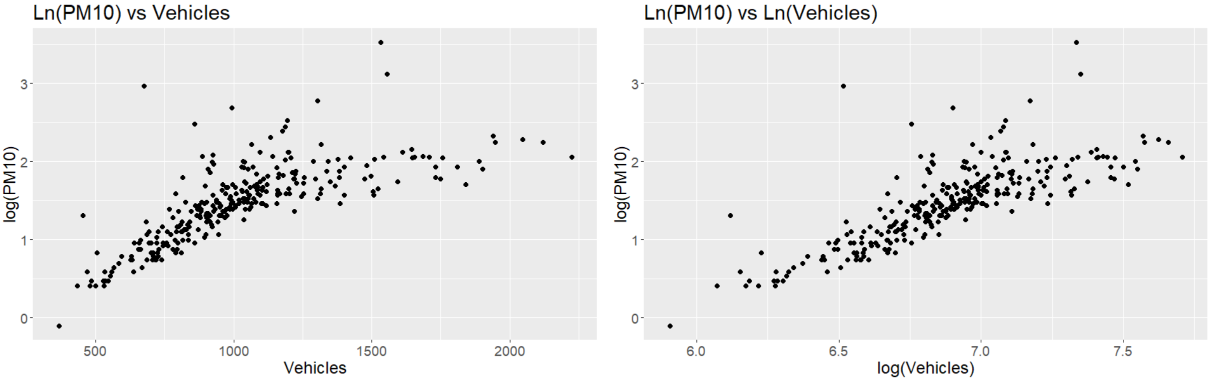


Figure 4: Residual plots for y=PM10(left) and y=ln(PM10)(right)

So I think picking x = ln(Vehicles) is a better option.

1.2 1(b)

1.2.1 Present Model:

The β-estimates are shown in Table 1 while 95 % confidence intervals are shown in Table 2.

Coefficient	**Estimate**
Intercept	-7.38912
log(Vehicles)	1.28693

Table 1: Regression Results

Coefficient	**2.5%**	**97.5%**
Intercept	-8.194491	-6.583741
log(Vehicles)	1.1.170085	1.403780

Table 2: 95 % confidence intervals

1.2.2 Plot:

The plot for ln(PM10) against ln(Vehicles) together with this estimated linear relationship, its 95 % confidence interval and a 95 % prediction interval for future observations is shown below.

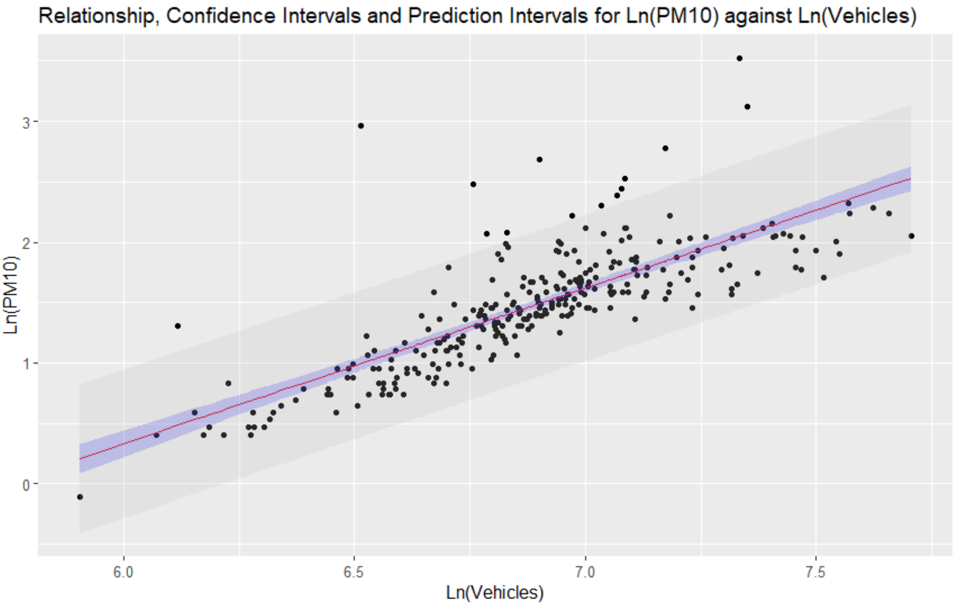


Figure 5: ln(PM10) against ln(Vehicles)

1.2.3 Transform back:

After transform, we have the model:

$$PM10 = e^{\beta_0 + \beta_1 ln(Vehicles)}$$

and the plot for this model is shown below.

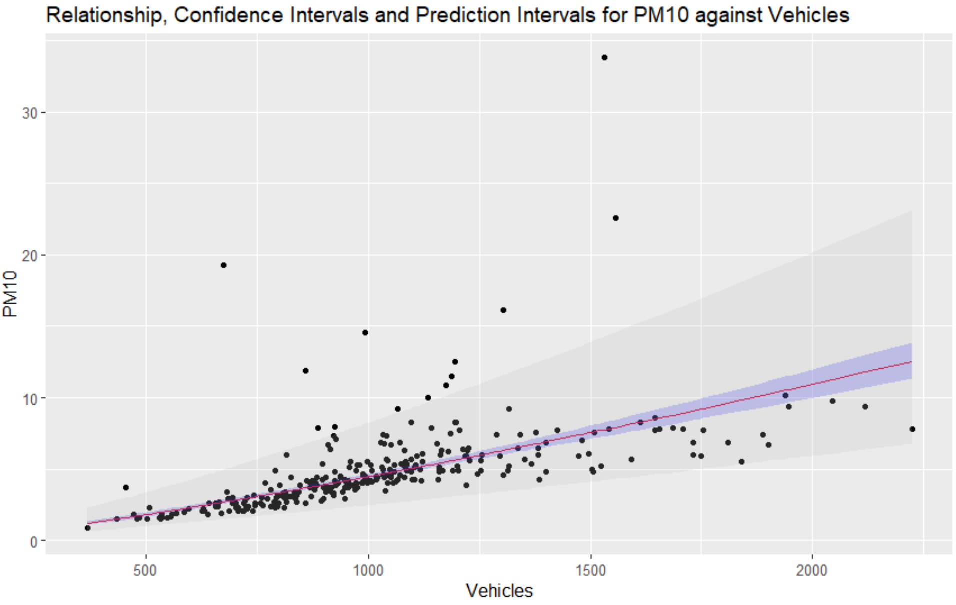


Figure 6: PM10 against Vehicles

We find that even the 95 % prediction interval does not cover all the data, indicating that there are still problems with the model fit, which is reflected in the residual plots as patterns (like a curve or systematic deviations).

1.3 1(c)

1.3.1 Decrease number of vehicles:

First, we calculate the logarithmic decrease:

$$\Delta ln(Vehicles) = ln(0.9 * Vehicles) - ln(Vehicles) = ln(0.9) \approx -0.10536$$

$$\Delta ln(PM10) = \beta_1 \Delta ln(Vehicles) \approx 1.28693 * (-0.10536) = -0.1356$$

So the change in PM10 is:

$$e^{\Delta ln(PM10) \approx e^{-0.1356} \approx 0.873}$$

which means PM10 will decrease to about 87.3% of the original value, a reduction of approximately 12.7%.
As the 95% confidence interval for β is $(1.17056, 1.4033)$, the confidence interval for the change in PM10 is $(e^{\beta_{1lower}*ln(Vehicles)}, e^{\beta_{1upper}*ln(Vehicles)})$, which is $(0.8838, 0.8626)$, meaning that PM10 will decrease to between 86.26% to 88.38% of its original value at 95% confidence.

1.3.2 Half PM10 emissions:

We can set up the equation:

$$ln(0.5 * PM10) = \beta_0 + \beta_1 ln(Vehicles_{New})$$

$$ln(0.5) = \beta_1(ln(Vehicles_{New}) - ln(Vehicles))$$

$$ln(Vehicles_{New}) = ln(Vehicles) + \frac{ln(0.5)}{\beta_1}$$

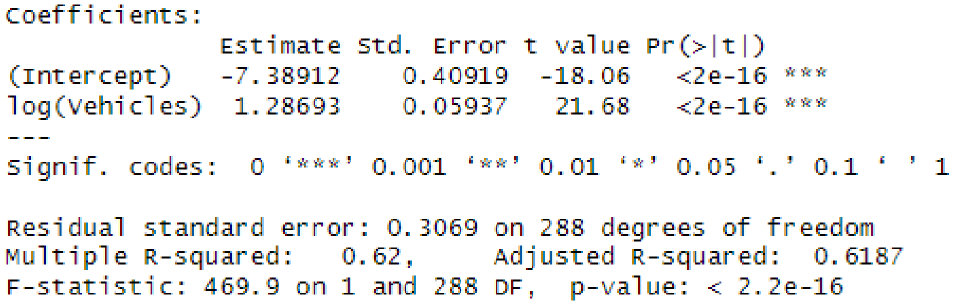
$$Vehicles_{New} = e^{ln(Vehicles) + \frac{ln(0.5)}{\beta_1}} \approx 0.583 * Vehicles$$

Thus, the number of vehicles must be reduced to about 58.3% of its original value to half PM10 emissions. Replacing β_1 with its upper and lower value at 95% confidence interval as the task above, we can get the interval for the reduction of vehicles is $(0.53, 0.61)$.

2 Adding more explanatory variables

2.1 2(a)

We think that there is a significant linear relationship between ln(PM10) and ln(Vehicles).Here is the explanation:
From the summary of model2(b), we can find the t-test value and P-value for β_1, which are 21.68 and $2e - 16$ respectively.
While the null hypothesis H_0 is $\beta_1 = 0$, which states that there is no linear relationship between the log of vehicles numbers and the log of PM10 emissions, the high t-value and extremely small P-value indicating strong evidence against the null hypothesis. So we reject the null hypothesis H_0.
The t-statistic follows a t-distribution under the null hypothesis. The degrees of freedom for this test are equal to $n - k$, where n is the number of observations and k is the number of predictors (including the intercept).



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -7.38912    0.40919  -18.06   <2e-16 ***
log(Vehicles)    1.28693    0.05937   21.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3069 on 288 degrees of freedom
Multiple R-squared:   0.62,	Adjusted R-squared:  0.6187 
F-statistic: 469.9 on 1 and 288 DF,  p-value: < 2.2e-16
```

Figure 7: Model1(b)

2.2 2(b)

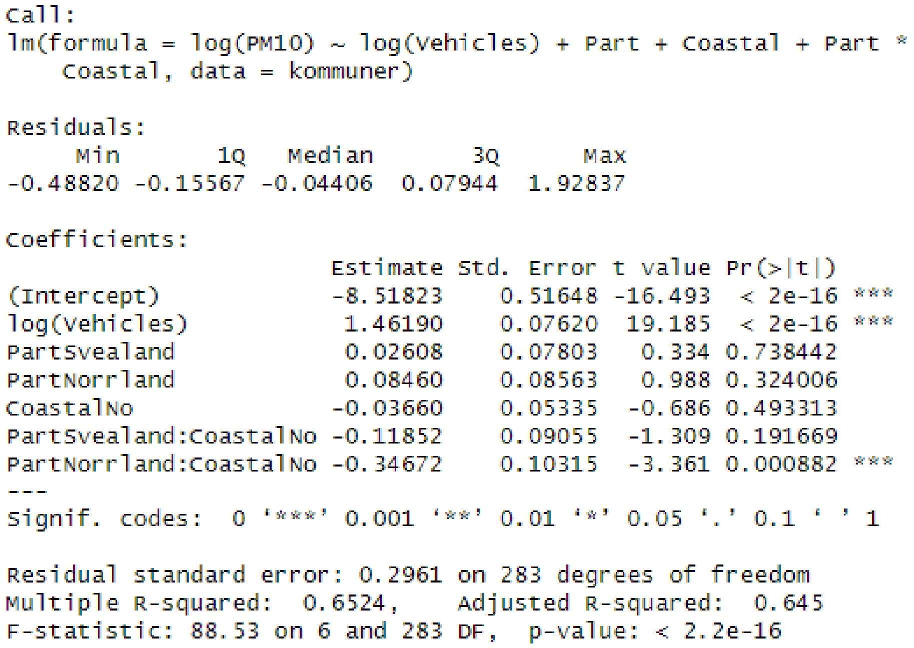
2.2.1 Numbers of 6 possible combinations:

Part	Coastal	number
Gotaland	No	92
Gotaland	Yes	48
Svealand	No	74
Svealand	Yes	22
Norrland	No	37
Norrland	Yes	17

Table 3: Part and Coastal Distribution

2.2.2 Model 2(b):

The new model, $Model2(b)$ is:



```
Call:
lm(formula = log(PM10) ~ log(Vehicles) + Part + Coastal + Part *
    Coastal, data = kommuner)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48820 -0.15567 -0.04406  0.07944  1.92837

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -8.51823    0.51648 -16.493  < 2e-16 ***
log(Vehicles)            1.46190    0.07620  19.185  < 2e-16 ***
PartSvealand             0.02608    0.07803   0.334 0.738442
PartNorrland             0.08460    0.08563   0.988 0.324006
CoastalNo               -0.03660    0.05335  -0.686 0.493313
PartSvealand:CoastalNo -0.11852    0.09055  -1.309 0.191669
PartNorrland:CoastalNo -0.34672    0.10315  -3.361 0.000882 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2961 on 283 degrees of freedom
Multiple R-squared:  0.6524,	Adjusted R-squared:  0.645
F-statistic: 88.53 on 6 and 283 DF,  p-value: < 2.2e-16
```

Figure 8: Model2(b)

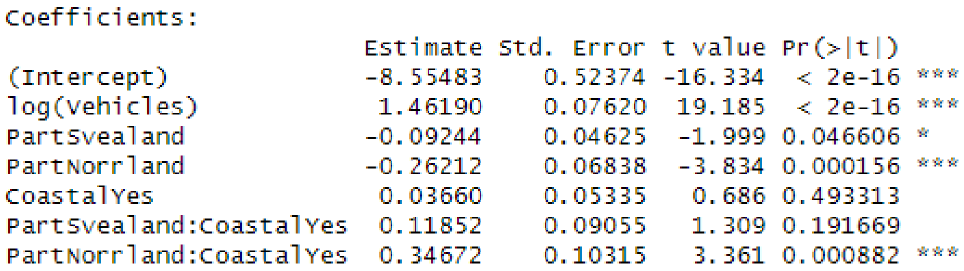
Coefficient	Estimate	95% CI
(Intercept)	-8.55483	(-9.58574790 -7.523904991)
log(Vehicles)	1.46190	(1.31191163 1.611896314)
PartSvealand	-0.09244	(-0.18347254 -0.001398484)
PartNorrland	-0.26212	(-0.39670851 -0.127529674)
CoastalYes	0.03660	(-0.06841879 0.141609089)
PartSvealand:CoastalYes	0.11852	(-0.05972928 0.296761225)
PartNorrland:CoastalYes	0.34672	(0.14368750 0.549746709)

Table 4: Regression Results and 95 % confidence intervals

The category here are "Gotaland" and "No" respectively, as they are the 1st level. It's a suitable reference as these two categories has the largest observation numbers.

2.2.3 Parameters test:

The t-test and P-value can be found in the summary of model2(b)



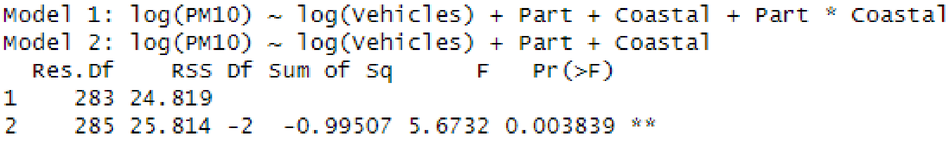
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -8.55483    0.52374 -16.334  < 2e-16 ***
log(Vehicles)              1.46190    0.07620  19.185  < 2e-16 ***
PartSvealand              -0.09244    0.04625  -1.999 0.046606 *  
PartNorrland              -0.26212    0.06838  -3.834 0.000156 ***
CoastalYes                 0.03660    0.05335   0.686 0.493313    
PartSvealand:CoastalYes    0.11852    0.09055   1.309 0.191669    
PartNorrland:CoastalYes    0.34672    0.10315   3.361 0.000882 ***
```

Figure 9: Model2(b) test

The t-statistic follows a t-distribution under the null hypothesis. The degrees of freedom for this test are equal to $n - k$, where n is the number of observations and k is the number of predictors (including the intercept).

The P-value for coefficient ”CoastalYes” and ”PartSvealand:CoastalYes” are large than 0.05, while the null hypothesis H_0 is $\beta = 0$, which states that there is no linear relationship between the coefficients and the log of PM10 emissions, the high P-value indicating week evidence against the null hypothesis. So we accept the null hypothesis H_0. The other coefficients are significantly different from zero.

2.2.4 Interaction test:



```
Model 1: log(PM10) ~ log(Vehicles) + Part + Coastal + Part * Coastal
Model 2: log(PM10) ~ log(Vehicles) + Part + Coastal
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)   
1    283 24.819                                
2    285 25.814 -2  -0.99507 5.6732 0.003839 **
```

Figure 10: Intercation Analysis

The P-value is 0.003839, which is much smaller than 0.05, indicates that the interaction is significant. The null hypothesis and other stuffs are similar to the last question.

2.2.5 log-PM10 for 6 part/coastal combinations:

Coastal	Part	Part	Lower CI	Upper CI
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 5: log-PM10 for 6 part/coastal combinations

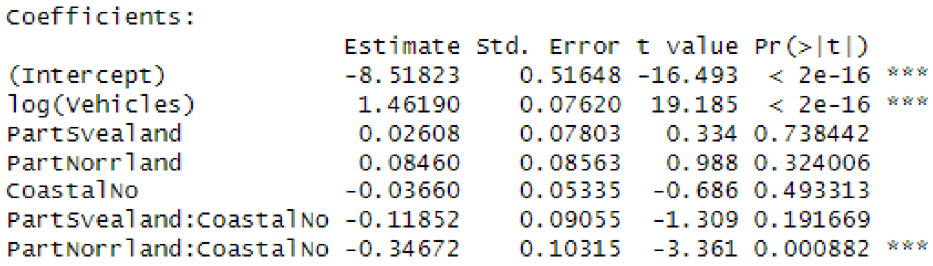
2.2.6 PM10 for 6 part/coastal combinations:

Coastal	Part	Prediction	Lower CI	Upper CI
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 6: PM10 for 6 part/coastal combinations

2.3 2(c)

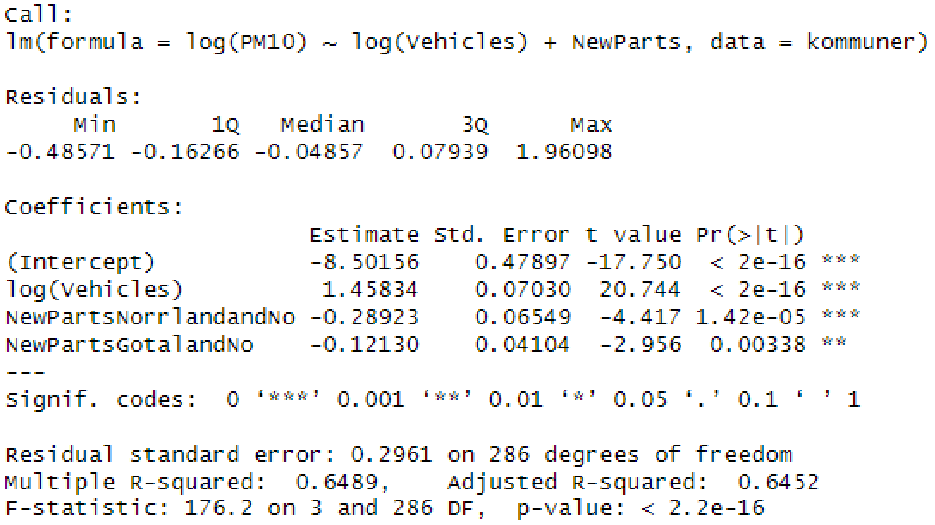
After using Coastal=”yes” as reference ,we can get the model:



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -8.51823    0.51648 -16.493  < 2e-16 ***
log(Vehicles)              1.46190    0.07620  19.185  < 2e-16 ***
PartSvealand               0.02608    0.07803   0.334 0.738442    
PartNorrland               0.08460    0.08563   0.988 0.324006    
CoastalNo                 -0.03660    0.05335  -0.686 0.493313    
PartSvealand:CoastalNo    -0.11852    0.09055  -1.309 0.191669    
PartNorrland:CoastalNo    -0.34672    0.10315  -3.361 0.000882 ***
```

Figure 11: model2(b) with reference Coastal=”Yes”

Comparing model2(b), we find that the P-values of ”PartSvealand” and ”PartNorrland” coefficients are larger than 0.05 when ”Yes” is used as the reference category, and we should accept the null hypothesis, which suggests that ”PartSvealand” and ”PartNorrland” do not show a significant difference when conditioned on Coastal=”Yes”. So we can choose ”coastal Svea/Norrland” as our first variable; and we also find that the ”PartSvealand:CoastalNo” and ”PartSvealand:CoastalYes” parameters both have large P-values, presumably because Gotaland’s Coastal or not as a reference does not result in a significant change. So the first variable can now be ”Gotaland coastal Svea/Norrland”, while the remaining analogies each act as a variable.
In the spirit of reducing the parameters while maintaining the model fit, we selected three variables as **1= Gotaland or coastal Svea/Norrland, 2 = Inland-Svealand, 3 = Inland-Norrland**, which only need 2 extra coefficients to fit, much less than the 5 coefficients needed originally.
The finnal model 2(c) we get is:



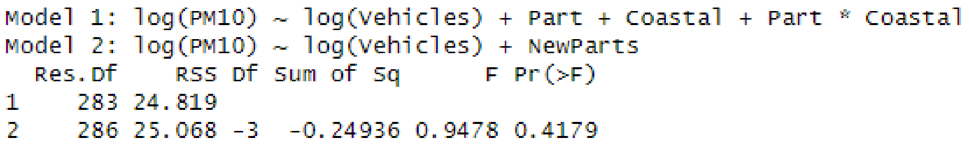
```
Call:
lm(formula = log(PM10) ~ log(Vehicles) + NewParts, data = kommuner)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.48571 -0.16266 -0.04857  0.07939  1.96098 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -8.50156    0.47897 -17.750  < 2e-16 ***
log(Vehicles)          1.45834    0.07030  20.744  < 2e-16 ***
NewPartsNorrlandandNo -0.28923    0.06549  -4.417 1.42e-05 ***
NewPartsGotalandNo    -0.12130    0.04104  -2.956  0.00338 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2961 on 286 degrees of freedom
Multiple R-squared:  0.6489,	Adjusted R-squared:  0.6452 
F-statistic: 176.2 on 3 and 286 DF,  p-value: < 2.2e-16
```

Figure 12: Model 2(c)

Variable	Estimate	95% CI
(Intercept)	-8.50156	(-9.4443220, -7.55880124)
log(Vehicles)	1.45834	(1.3199682, 1.59671687)
NewPartsNorrlandandNo	-0.28923	(-0.4181248, -0.16033497)
NewPartsGotalandNo	-0.12130	(-0.2020737, -0.04053269)

Table 7: Regression Results and 95 % confidence intervals

We can compare model2(c) with model2(b) with command ”anova()” to see the difference:



```
Model 1: log(PM10) ~ log(Vehicles) + Part + Coastal + Part * Coastal
Model 2: log(PM10) ~ log(Vehicles) + NewParts
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    283 24.819
2    286 25.068 -3  -0.24936 0.9478 0.4179
```

Figure 13: Difference between model2(b) and model2(c)

As the P-value is 0.4179, which is much larger than 0.05, suggesting that the two models are not very different, while model2(c) uses only two additional parameters, suggesting that the selection of the new variables is effective.

2.4 2(d)

2.4.1 Plot log-PM10 against each of numerical variables:

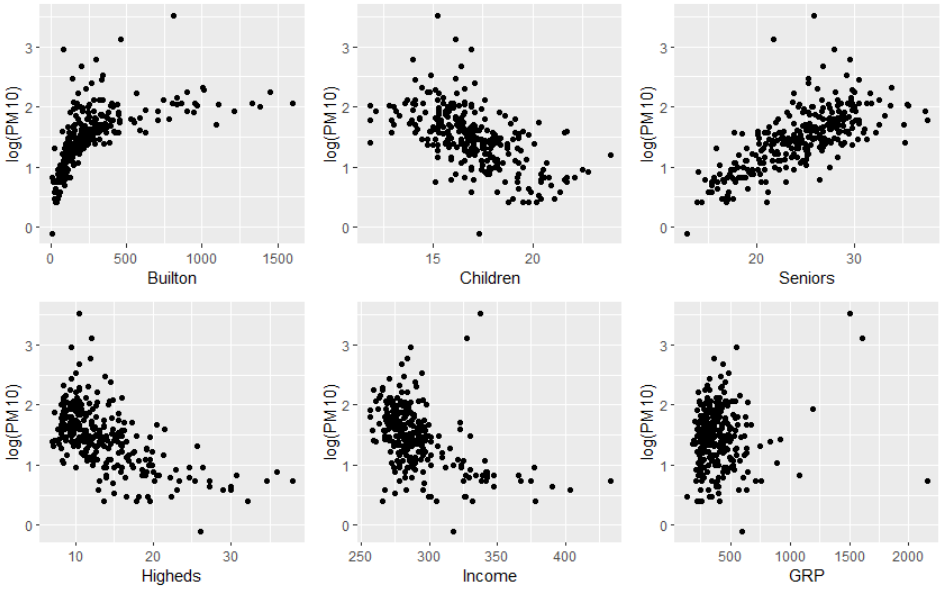


Figure 14: log-PM10 against each of numerical variables

We can notice that "Builton", "Higheds" and "Income" show a log-relationship with log-PM10.The plots after transform is:

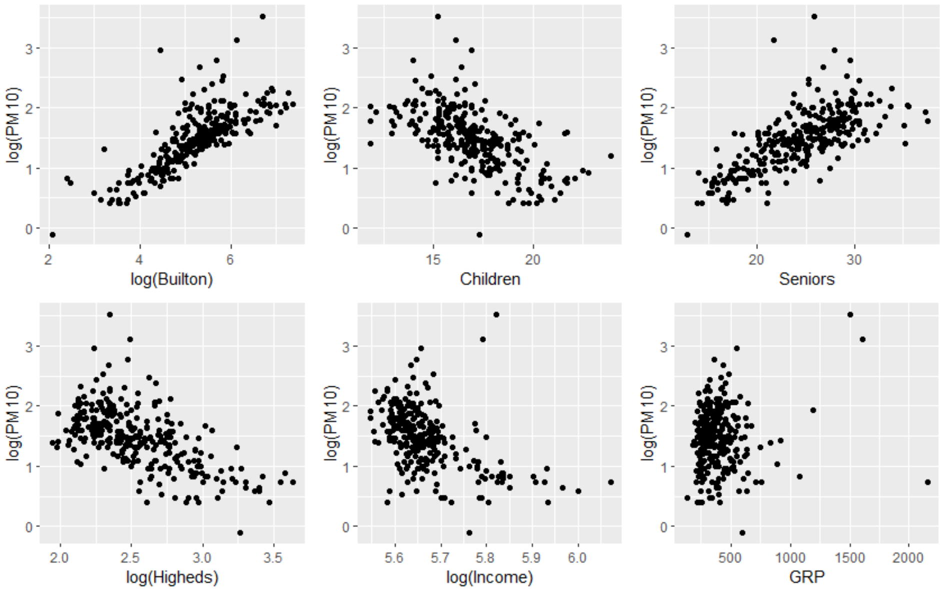
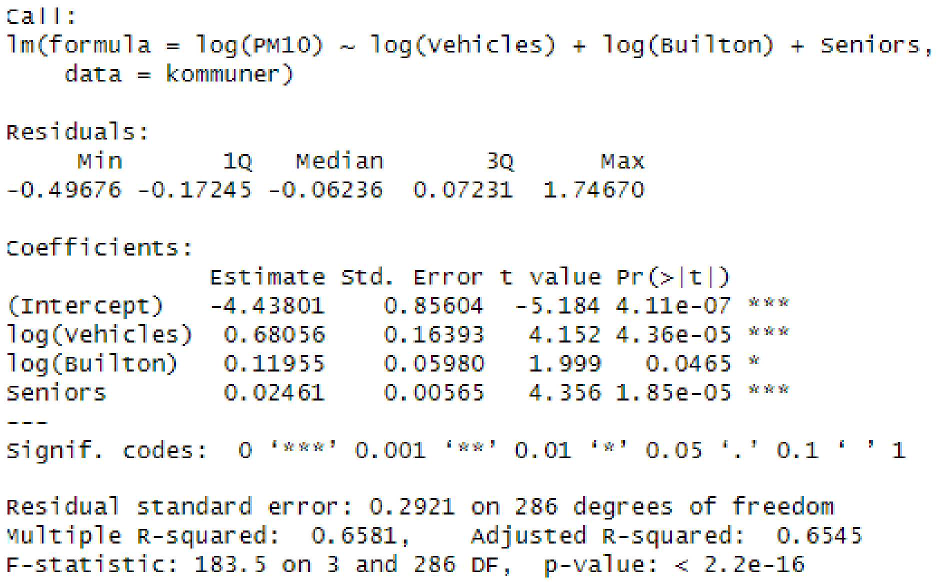


Figure 15: log-PM10 against each of numerical variables(log-transformed)

2.4.2 Fit a model with three explanatory variables:

Based on plots we chose log(Builton) and log(Seniors) as variables strongly correlated with log-PM10.The model fitted with log(Vehicles) + log(Builton) + Seniors is shown below:



```
Call:
lm(formula = log(PM10) ~ log(Vehicles) + log(Builton) + Seniors,
    data = kommuner)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.49676 -0.17245 -0.06236  0.07231  1.74670 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -4.43801    0.85604  -5.184 4.11e-07 ***
log(Vehicles)  0.68056    0.16393   4.152 4.36e-05 ***
log(Builton)   0.11955    0.05980   1.999   0.0465 *  
Seniors        0.02461    0.00565   4.356 1.85e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2921 on 286 degrees of freedom
Multiple R-squared:  0.6581,	Adjusted R-squared:  0.6545 
F-statistic: 183.5 on 3 and 286 DF,  p-value: < 2.2e-16
```

Figure 16: Model with three explanatory variables

The β -estimates, their standard errors, confidence intervals and the P-values for their t-tests are shown below:

Variable	Estimate	Std. Error	95% CI	P-value
(Intercept)	-4.43801	0.85604	(-6.122950495, -2.75307344)	4.11e-07
log(Vehicles)	0.68056	0.16393	(0.357905506, 1.00321300)	4.36e-05
log(Builton)	0.11955	0.05980	(0.001855236, 0.23725126)	0.0465
Seniors	0.02461	0.00565	(0.013489049, 0.03573224)	1.85e-05

Table 8: Values for Model

To determine possible multicollinearity problems, we first plot a scatter plot between each pair of variables to see the linear relationship between them:

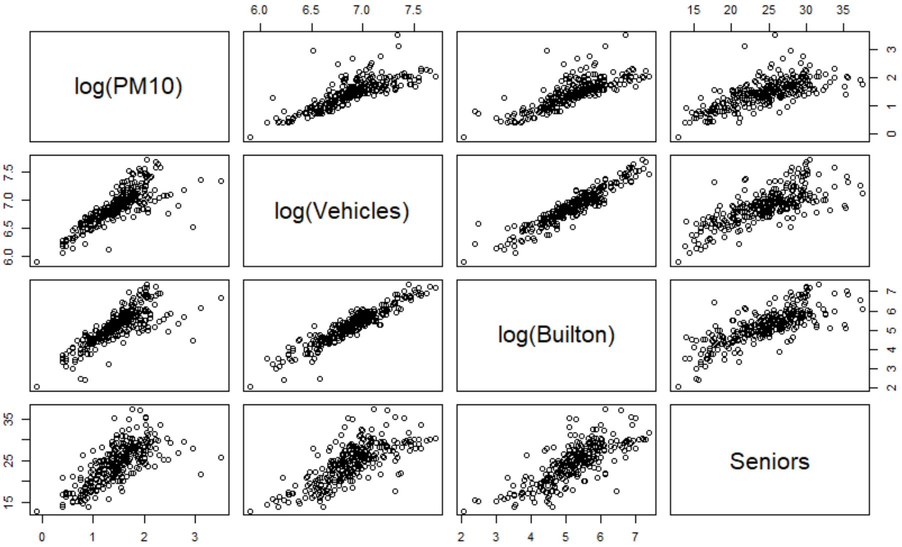


Figure 17: Scatter plot between each pair of variables

It is clear from the figure that we can find a strong covariance between log(Vehicles) and log(Builton).This is explicable because places with more cars are indicative of a developed industry and economy, and with that comes a wider coverage of buildings and roads.
We can also assess the degree of multicollinearity by calculating the correlation coefficients between the variables and calculating the VIF for each explanatory variable:

	log_PM10	**log_Vehicles**	**log_Builton**	**Seniors**
log_PM10	1.0000000	0.7874094	0.7818022	0.7030519
log_Vehicles	0.7874094	1.0000000	0.9383257	0.7358597
log_Builton	0.7818022	0.9383257	1.0000000	0.7658717
Seniors	0.7030519	0.7358597	0.7658717	1.0000000

Table 9: Correlation matrix for log-transformed variables

	log(Vehicles)	**log(Builton)**	**log(Seniors)**
VIF	8.415566	9.332962	2.433331

Table 10: VIF values for three variables

Both the matrix of correlation coefficients(0.9383257) and the VIF values(> 5) also demonstrate again the covariance of log(Vehicles) and log(Builton).

2.4.3 Exclude one variable:

Obviously log(Builton) should be excluded. Refit the model to get Model 2(d), β-estimates and their standard errors, confidence intervals and P-values are shown below:

Variable	**Estimate**	**Std. Error**	**95% CI**	**P-value**
(Intercept)	-5.845624	0.489491	(-6.80907092, -4.88217660)	2e-16
log(Vehicles)	0.962652	0.083885	(0.79754286, 1.12776030)	2e-16
Seniors	0.028249	0.005377	(0.01766487, 0.03883246)	2.92e-07

Table 11: Values for Model 2(d)

And the new VIF-values is 2.180975, which is acceptable.

2.4.4 Changes in standard errors:

In the presence of multicollinearity, high correlation between the independent variables leads to increased uncertainty in the estimation of the model parameters, which makes the standard errors of the parameter estimates increase. That's why we find the standard errors decrease in the second model(from 0.16319 to 0.08538) which excludes the multicollinearity variable, while VIF-value decreases at the same time.

2.5 2(e)

Fit a new model using the new categorisation from 2(c) and all the continuous variables, with GVIF values as shown in Table 12.

Generalized Variance Inflation Factor (GVIF) is an extension of the VIF used to address situations involving categorical predictor variables (factor variables). When the model includes multi-level categorical variables, traditional VIF calculations may not be applicable or sufficient to provide accurate measures of multicollinearity.

In our case, since we introduced NewParts obtained from model 2c into the new model, which is a multi-level categorical variable, the VIF values are now referred to as GVIF values. This adjustment allows for a more comprehensive assessment of multicollinearity in models containing both continuous and categorical variables.

Builton is the most problematic variable, with exceptionally high GVIF values (12.049992). Builton represents the 'Area covered in buildings, roads, etc (= not nature), (hectares / 1000 inhabitants)', indicating the urban area's size and level of development. We suspect its issues stem from strong linear correlations with multiple variables, such as log(Vehicles) and GRP. This is because a highly developed city often implies more vehicles and higher income levels.

Seniors also present issues (GVIF=7.424164), as Seniors represent the percentage of individuals aged 65 and above. Seniors evidently exhibit a strong negative correlation with the Children variable (percentage of individuals aged 0-14 years).

By removing the Builton and Seniors variables, fitting the model model_2e, and testing the GVIF, the results in Table 13 demonstrate that the GVIF values for each variable in the filtered model are acceptable.

Table 12: Model's GVIF Before Remove

Variable	GVIF	GVIF^(1/(2*Df))
log(Vehicles)	9.763394	3.124643
log(Higheds)	3.085264	1.756492
Children	5.380864	2.319669
Seniors	7.424164	2.724732
log(Income)	2.398015	1.548552
log(GRP)	1.475519	1.214709
log(Builton)	12.049992	3.471310
NewParts	1.767139	1.152969

Table 13: Model_2e's GVIF

Variable	GVIF	GVIF^(1/(2*Df))
log(Vehicles)	3.095087	1.759286
log(Higheds)	2.737429	1.654518
Children	1.891784	1.375422
log(Income)	2.216976	1.488951
log(GRP)	1.164690	1.079208
NewParts	1.695319	1.141071

3 Model validation and selection

3.1 3(a) Leverage.

The definition of leverage effect is crucial in linear regression analysis, as it delineates the extent to which data points influence parameter estimates. Leverage effect is defined as the diagonal elements of the projection matrix (also known as the hat matrix), indicating how far an observation deviates from the centroid of the independent variables. High leverage points suggest that these observations may disproportionately impact the regression line, potentially distorting model predictions. Identifying these points is vital for model diagnostics, as they may represent outliers or influential observations requiring further examination.

By computing leverage values, we obtained the leverage-yhat distribution plot as depicted in the Figure 18. Reference lines at 1/n and 2*p+1/n were added for comparison.

Subsequently, we sorted the leverage values for all data points based on Model 2(e) and identified the six groups with the highest leverage, as shown in Table 14.

To investigate why these data points exhibit such high leverage values, we plotted the variables in the model and log(PM10) separately, categorizing them into three subplots based on NewParts classification in Figure 19. In these plots, we highlighted the data points with high leverage values and annotated their respective Kommun.

Upon observing these plots, we found that in the "NorrlandandNo" part, Gllivare and Kiruna were outliers in each plot. Similarly, Sundbyberg in "SvealandandNo" also exhibited outlier behavior. The presence of these outliers can significantly impact the regression line.

For Danderyd, Solna, and Lund, there were no apparent anomalies in the distribution. However, in the plot of log(Higheds), these three data points exhibited similar distributions, as did log(Vehicles). Considering the multicollinearity between log(Higheds) and log(Vehicles), this may contribute to the increased leverage values of the data.

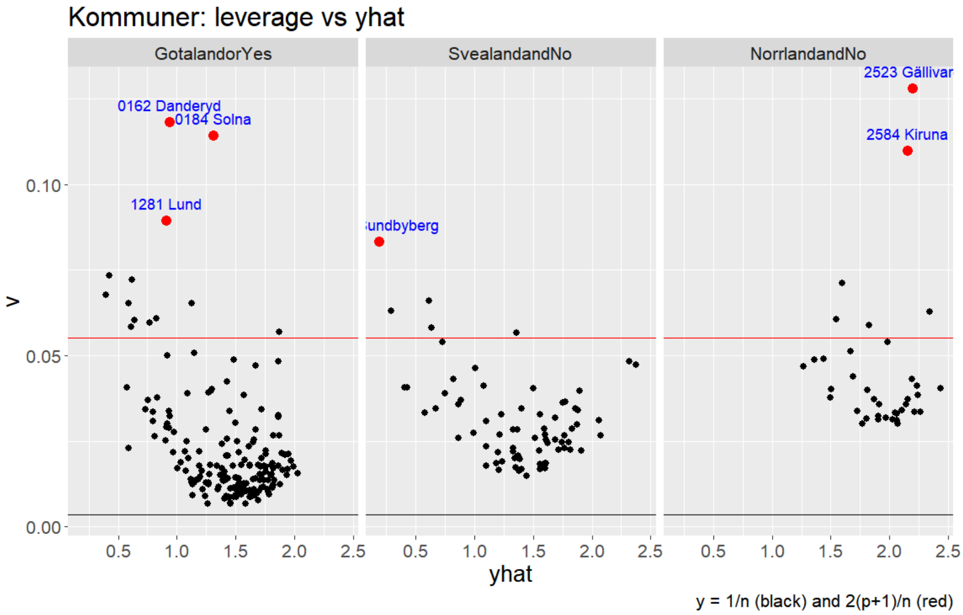


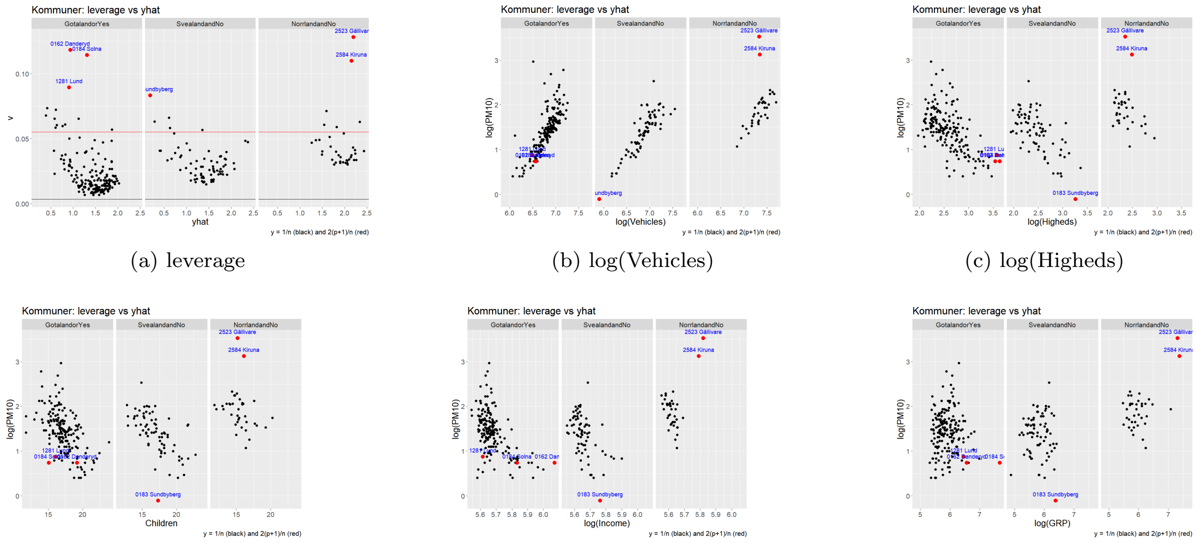
Figure 18: Leverage

3.2 3(b) Cooks distance.

Cook's distance is a powerful metric that combines leverage and prediction error information, providing a comprehensive measure of the influence of an observation. By quantifying the change in the entire model fit when a particular observation is excluded, Cook's distance can assess the impact of individual data points on the model. Observations with high Cook's distance suggest that they may play a decisive role in the regression model, warranting further investigation into whether they should be included in the dataset.

Table 14: 3-a Top Leverage

Kommun	**leverage**
2523 Gllivare	0.12812865
0162 Danderyd	0.11809561
0184 Solna	0.11417777
2584 Kiruna	0.10995045
1281 Lund	0.08930893
0183 Sundbyberg	0.08322614



(a) leverage

(b) log(Vehicles)

(c) log(Higheds)

(d) Children

(e) log(Income)

(f) log(GRP)

Figure 19: HighLight top Leverage

By computing Cook's distance, we obtained the Cook's distanceyhat distribution plot as shown in Figure 20, with reference lines added for 4/n (dashed), F_0.5, p+1, and n-(p+1) (solid).

Subsequently, we sorted the Cook's distance values for all data points based on Model 2(e) and identified the six groups with the highest Cook's distance, as shown in Table 15.

Some municipalities differ in terms of high leverage, as Leverage assesses the influence of observations on parameter estimates, while Cook's distance assesses the influence of observations on model fit. Cook's distance considers a broader scope and can identify outlier points that have a larger impact on model fit. However, Leverage is also an important metric that can identify observations that may be outliers.

Next, we investigated DFBETAS. DFBETAS, representing ”difference in coefficients,” refers to the difference in coefficient estimates when including and excluding a particular observation. Each DFBETAS value is standardized, providing a standardized measure of the influence of an observation on individual regression coefficients. Large DFBETAS values may indicate that the corresponding observation is an outlier for a particular predictor variable, aiding in the identification and handling of influential data points.

We plotted log-PM10 against the corresponding variable(s), highlighting the influential municipalities, and compared them with the DFBETAS values corresponding to the -parameters of log(PM10) against the respective variable(s), as shown in Figure 21. Upon observation, we found that data points with high DFBETAS and Cook's distance values tend to be outliers, distributed far from other data points, suggesting they may be anomalous data.

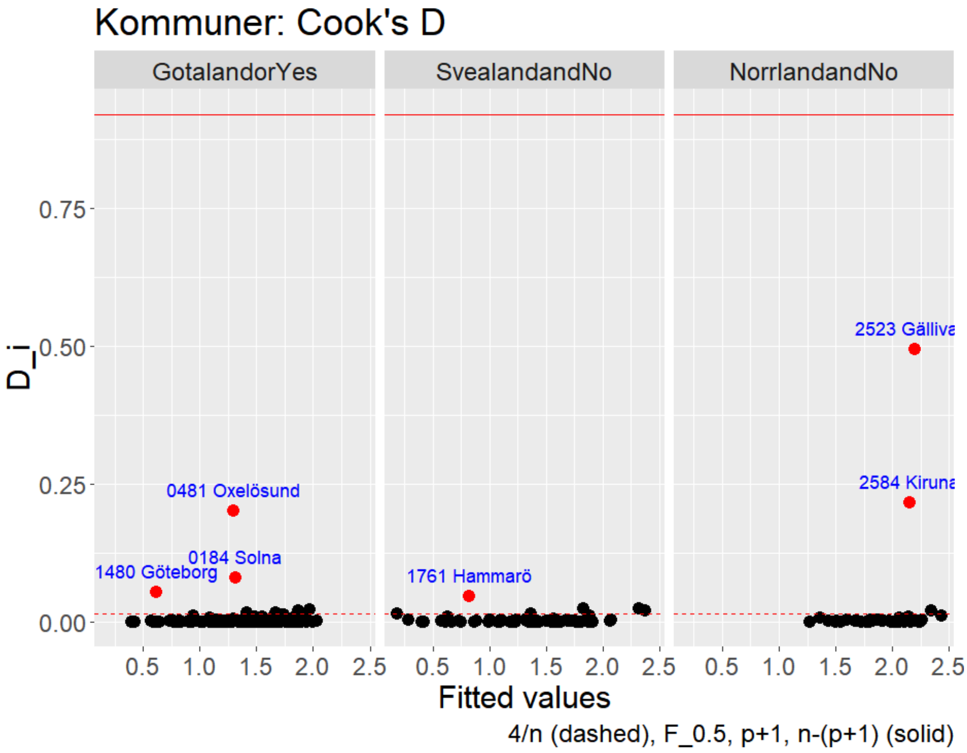
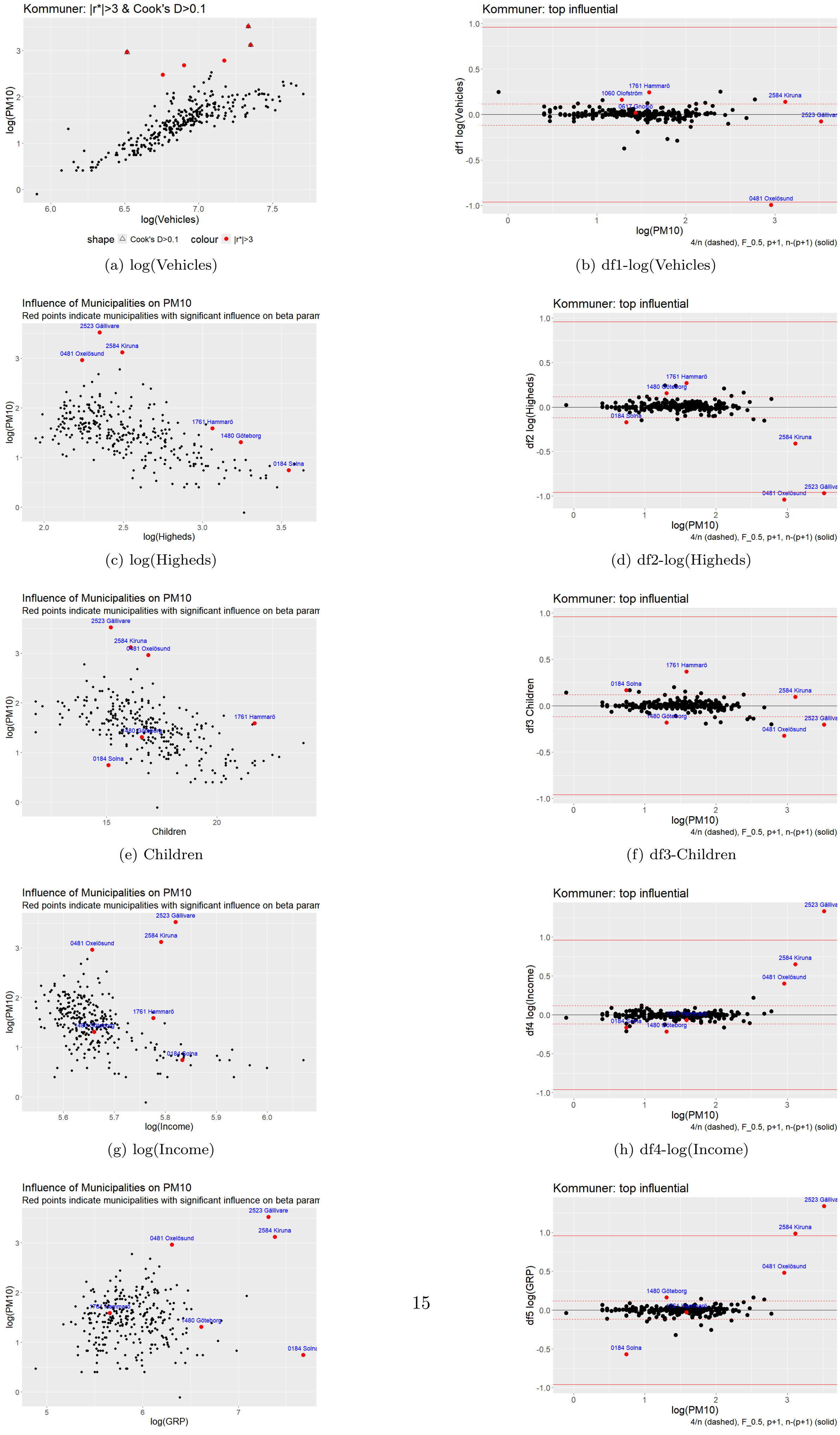


Figure 20: Cook's Distance

3.3 3(c) Studentized residuals.

Standardized residuals are a form of residuals used in regression analysis to identify outliers. Unlike simple residuals, standardized residuals are scaled based on the estimated standard deviation, which varies for different observations. This standardization process allows all residuals to be assessed using a single threshold to identify outliers. Standardized residuals with absolute values greater than 3 or less than -3 are typically considered outliers, requiring further investigation to determine if they exhibit leverage effects or influence the model.

By computing Studentized residuals, we obtained the Studentized residuals-yhat distribution plot as shown in Figure 23, with reference lines added for 4/n (dashed), F_0.5, p+1, and n-(p+1) (solid).



(a) log(Vehicles)

(b) df1-log(Vehicles)

(c) log(Higheds)

(d) df2-log(Higheds)

(e) Children

(f) df3-Children

(g) log(Income)

(h) df4-log(Income)

(i) log(GRP)

(j) df5-log(GRP)

Table 15: 3-b Highest Cooks distance municipalities

Kommun	**D**
2523 Gllivare	0.49362365
2584 Kiruna	0.21660607
0481 Oxelsund	0.20140791
0184 Solna	0.08028626
1480 Gteborg	0.05338203
1761 Hammar	0.04644829

Next, we highlighted the Studentized residuals values for the top Cook's distance data. Their residuals are presented in Table 16.

Subsequently, we identified all municipalities with $|ri| > 3$, as shown in Table 17, where Kalix, Karlshamn, and Mnsters did not have high Cook's distance values.

Similarly, we plotted $|ri|$ against the linear predictor, as shown in Figure 22, with reference lines at y = sqrt(0.75 quantile of normal), sqrt(2), and sqrt(3). The variance appears not to be constant.

Table 16: 3-c Highest Cooks distance municipalities' residual

Kommun	**r**
2523 Gllivare	5.440261
2584 Kiruna	3.835278
0481 Oxelsund	6.675051
0184 Solna	-2.248268
1480 Gteborg	2.649813
1761 Hammar	2.906809

Table 17: 3-c Highest residual municipalities

Kommun	**—abs(r)**
0481 Oxelsund	6.675051
0861 Mnsters	3.823718
1082 Karlshamn	3.994756
2514 Kalix	3.032695
2523 Gllivare	5.440261
2584 Kiruna	3.835278

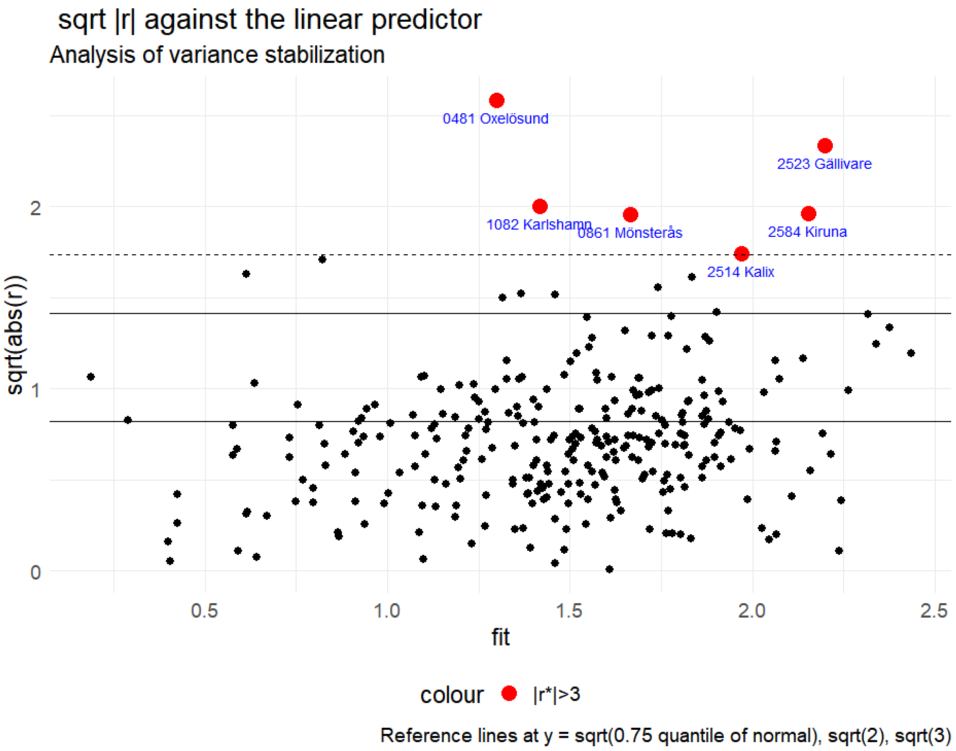


Figure 22: sqrt(r) Plot)

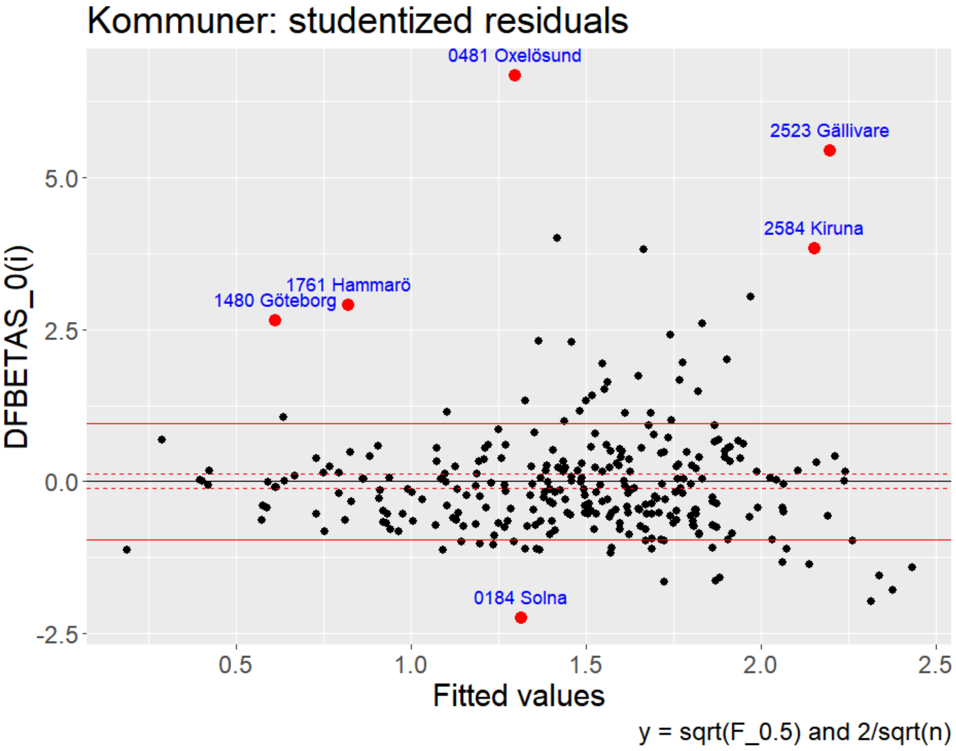


Figure 23: Top Cook's Distance Residuals

3.4 3(d) Explain, exclude, refit.

Based on the information provided, we understand that the main source of PM10 is domestic transportation, accounting for 40% of Sweden's emissions, which explains the high correlation between the number of vehicles and PM10. The second largest source of PM10 emissions in Sweden is industry, accounting for 32% of emissions. This aspect is not reflected in the data features, and our focus is more on per capita PM10 emissions rather than whether a small municipality happens to have a large PM10-emitting company. Therefore, we need to remove some data points.

Based on the table and the plot of sqrt(—r*—) vs. refitted values, the following municipalities are to be removed: ”0481 Oxelsund”, ”1082 Karlshamn”, ”0861 Mnsters”, ”2523 Gllivare”, ”2514 Kalix”, ”2584 Kiruna”, ”1480 Gteborg”, ”1761 Hammar”, ”1484 Lysekil”, ”1494 Lidkping”, ”1882 Askersund”, ”2284 rnskldsvik”, ”0319 lvkarleby”, ”1460 Bengtsfors”, ”1781 Kristinehamn”, ”2262 Timr”, ”0980 Gotland”, ”1272 Bromlla”, ”1885 Lindesberg”, and ”1764 Grums”. This filtering results in the refitted model, model_3d. Plotting sqrt(—r*—) vs. refitted values shows the distribution as depicted in Figure 24.

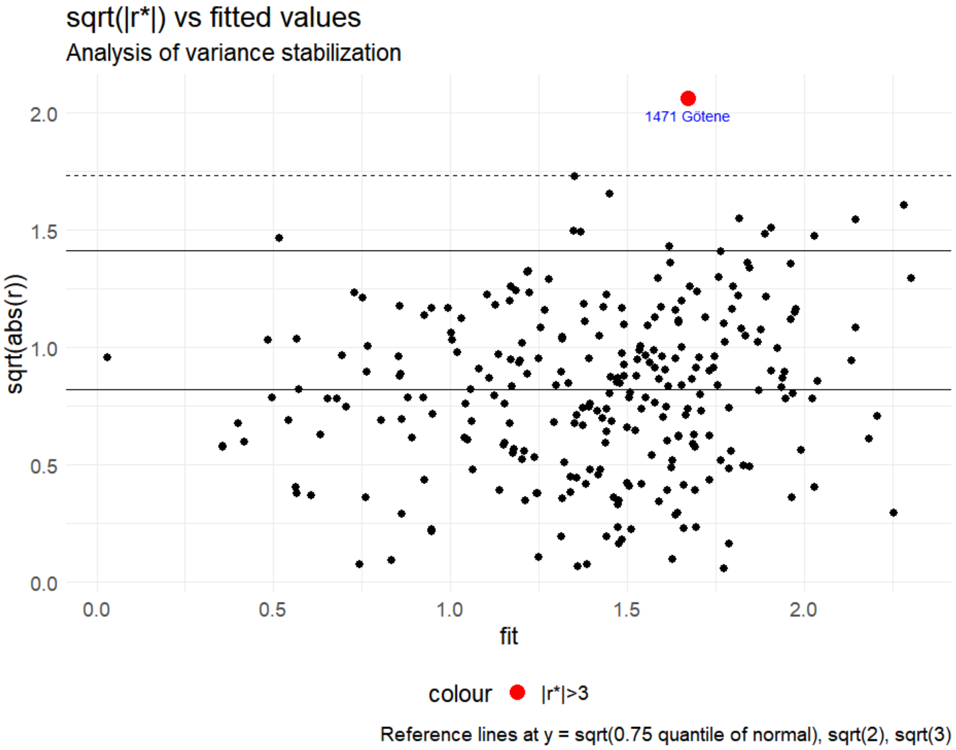


Figure 24: Refitted $|r|$

We then compare model_3d (with Confidence Intervals as shown in Table 18 and test statistics as shown in Table 20) with model_2e (with Confidence Intervals as shown in Table 19 and test statistics as shown in Table 21) to assess how well the assumptions of normality and constant variance of the residuals hold. By observing the Q-Q plot and the residual-fitted values plot, as shown in Figure 25, we find that the residuals of model_3d generally follow a normal distribution, while model_2e exhibits some deviations and does not fully meet the assumption of normality.

Table 18: Model 3d Confidence Intervals

Variable	**2.5%**	**97.5%**
(Intercept)	-4.419	-0.307
log(Vehicles)	1.097	1.311
log(Higheds)	-0.202	-0.015
Children	-0.030	-0.006
log(Income)	-1.047	-0.323
log(GRP)	-0.054	0.061
NewPartsSvealandandNo	-0.167	-0.079
NewPartsNorrlandandNo	-0.346	-0.203

Table 19: Model 2e Confidence Intervals

Variable	**2.5%**	**97.5%**
(Intercept)	-10.073	-3.326
log(Vehicles)	0.936	1.302
log(Higheds)	-0.466	-0.155
Children	-0.054	-0.011
log(Income)	-0.479	0.725
log(GRP)	0.106	0.292
NewPartsSvealandandNo	-0.202	-0.051
NewPartsNorrlandandNo	-0.381	-0.135

Table 20: Model 3d Coefficients

Variable	**Estimate**	**Std. Error**	**t value**	$Pr(>\|t\|)$
(Intercept)	-2.363	1.044	-2.263	0.024
log(Vehicles)	1.204	0.054	22.194	$< 2e - 16$
log(Higheds)	-0.108	0.048	-2.282	0.023
Children	-0.018	0.006	-2.948	0.003
log(Income)	-0.685	0.184	-3.723	0.000
log(GRP)	0.004	0.029	0.121	0.904
NewPartsSvealandandNo	-0.123	0.022	-5.513	0.000
NewPartsNorrlandandNo	-0.274	0.036	-7.595	0.000

Table 21: Model 2e Coefficients

Variable	**Estimate**	**Std. Error**	**t value**	$Pr(>\|t\|)$
(Intercept)	-6.699	1.714	-3.909	0.000
log(Vehicles)	1.119	0.093	12.044	$< 2e - 16$
log(Higheds)	-0.311	0.079	-3.932	0.000
Children	-0.032	0.011	-3.027	0.003
log(Income)	0.123	0.306	0.402	0.688
log(GRP)	0.199	0.047	4.198	¡ 0.001
NewPartsSvealandandNo	-0.126	0.038	-3.294	0.001
NewPartsNorrlandandNo	-0.258	0.063	-4.118	¡ 0.001

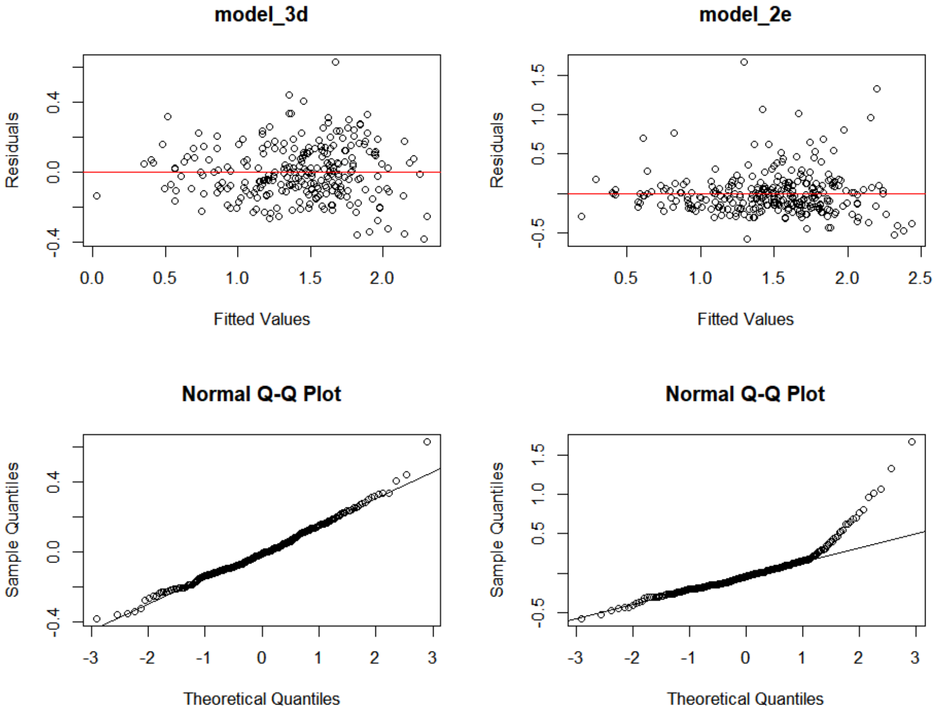


Figure 25: Residual Distribution Comparison

3.5 3(e) Variable selection.

Table 22 presents a comparison of the performance of six different models, including the number of -parameters, residual standard deviation, R-squared, adjusted R-squared, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Below is a detailed explanation of each metric and how they can be used to assess model performance:

1. **Metric Explanation:**

   Beta_Parameters (Number of Model Parameters): Indicates the number of independent variables in the model. More parameters may imply a more complex model.

   Residual_SD (Residual Standard Deviation): Measures the standard deviation of model residuals, reflecting the average magnitude of the differences between predicted and observed values. A smaller residual standard deviation generally indicates higher predictive accuracy of the model.

   R_Squared: Measures the proportion of total variability explained by the model. The closer the R-squared value is to 1, the greater the variability explained, indicating a better model fit.

   Adjusted_R_Squared: An R-squared adjusted for the number of model parameters. It adjusts for degrees of freedom, decreasing when the model includes insignificant variables, making it a better metric for comparing models with different numbers of independent variables.

   AIC (Akaike Information Criterion): Assesses the overall quality of the model considering both the complexity of the model and its goodness of fit. Lower AIC values indicate a preferable model.

   BIC (Bayesian Information Criterion): Similar to AIC but imposes a higher penalty for the number of parameters, suitable for larger sample sizes. Lower BIC values indicate a preferable model. Assessing Model Performance:

2. **Selecting the Optimal Model:**

   Observe Adjusted R-Squared: The highest value generally indicates the best model fit.

   Observe Residual_SD: The lowest value indicates the smallest prediction error.

   AIC and BIC: The model with the lowest values is typically the best, as these criteria consider both model complexity and fit.

3. **Conclusion:**

   From the provided data, Model 3(d) and the AIC Model perform well in terms of Adjusted R-Squared, AIC, and BIC, with the AIC Model having the lowest AIC value, suggesting a good fit while considering model complexity. If the sample size is sufficient, BIC is also an important consideration, and here too, the AIC Model performs well.

   These metrics collectively assist in evaluating and selecting models from a statistical perspective, but the final choice should also consider practical application needs, such as model interpretability, computational cost, and prediction accuracy.

Table 22: Comparison of Model Performances

Model	**Beta**	**Residual_SD**	**R_Squared**	**Adjusted_R_Squared**	**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

4 Team Roles

Yifei Zhang is responsible for problem analysis, code writing, and drafting report part 3.
Jialu Xu is responsible for problem analysis, code writing and review, and drafting report parts 1 & 2.

5 AI Usage Statement

In this project, AI is utilized solely for the following purposes:
1. Translation from Chinese to English.
2. Generating tables.
3. Consultation and resolution of certain code-related issues.