Errata List for Linear and Logistic Regression - Project 1

Kuan Teh Wan,Wanru Cheng

May 2024

1 2(a)

To calculate the degrees of freedom for the t-distribution, we have that (t(n-(p+1))), for **n = total number of observation** and **p = the number of variables**. Thus the degrees of freedom for this t-test is 288, t(290-(1+1))=t(288).

2 2(c)

To further reduce to the number of necessary combinations, we create a new variable based on this.

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

And by doing Partial F-test to compare this 4-category model with the original 6-cat, we have that the P-value is larger than 0.05, therefore, not being significantly different from the original.

Its estimates and confidence intervals in the table below.

	Estimates	95% C.I
Yes/Gotaland	-8.3772459	(-9.33483106, -7.41966077)
Yes/Svealand	-8.3772459	(-9.33483106, -7.41966077)
No/Gotaland	-8.3772459	(-9.33483106, -7.41966077)
No/svealand	-0.1102790	(-0.19241863, -0.02813947)
Yes/Norrland	0.1068103	(-0.04452828, 0.25814883)
No/Norrland	-0.2684597	(-0.40046963, -0.13644972)

Table 1: Expected log(PM10) and 95% confidence intervals for *Model.2(c)*

3 2(d)

$\sqrt{VIF_i}$ tells us how many times larger the standard error is due to the dependence on the other variables. Therefore, we see that the log-Vehicles standard error decreased from 0.16319 to 0.08538(halved) when the VIF decrease by a factor of 4 (8.458155 $\rightarrow$ 2.293231). Thus, the new model is the better choice.

4 2(e)

Now we'll come up with a new model with all the continuous variables except for Builton as discussed previously. The new model as follow:

$$\begin{aligned} ln(PM10) = \beta_0 + \beta_1 ln(Vehicles) + \beta_2 ln(Higheds) + \beta_3 ln(Seniors) \\ + \beta_4 Children + \beta_5 ln(Income) + \beta_6 ln(GRP) + \beta_7 NewParts2 + \epsilon \end{aligned} \tag{2}$$

The VIF values are now GVIF values since we have included categorical variable in our model. Next, we inspect these GVIF values to determine whether is it safe to use all the variables in the same model. Values are presented in the table below. The plot of the variables' correlations against all the others is also shown.

	GVIF value
log(Vehicles)	4.213032
log(Higheds)	3.078939
log(Seniors)	6.414595
Children	4.502060
log(Income)	2.346658
log(GRP)	1.421348
NewParts2	1.988422

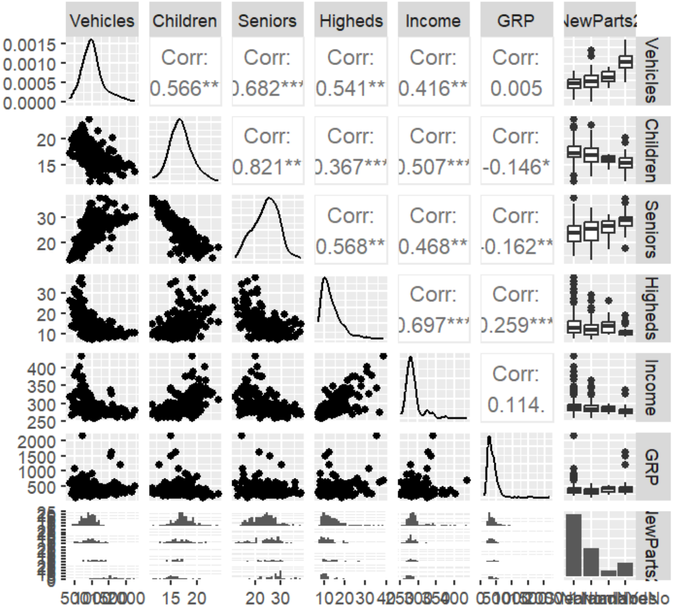

Figure 1: correlation

From the GVIF table, we can see that variable Senior has the highest GVIF value when compared to others. On the other hand, from the correlation plot we can also see that variable Senior has high correlation with other variables, especially with variable Children, correlation= -0.821.

To sum up, from both GVIF value and correlation, it suggests that variable Senior is the most problematic of all, therefore, must be excluded from the model.Thus we the Model2(e):

$$\begin{aligned} ln(PM10) &= \beta_0 + \beta_1 ln(Vehicles) + \beta_2 ln(Higheds) \\ +\beta_3 Children &+ \beta_4 ln(Income) + \beta_5 ln(GRP) + \beta_6 NewParts2 + \epsilon \end{aligned} \tag{3}$$

The new GVIF values for Model2(e) are listed in the table:

	GVIF value
log(Vehicles)	3.246481
log(Higheds)	2.852111
Children	1.893269
log(Income)	2.251550
log(GRP)	1.164923
NewParts2	1.854209

Part 3. Model validation and selection

3.(a) Leverage

From the model-2e we gainded the the last section, we are here to comduct the model validation and selection.

$$\begin{aligned} ln(PM10) = \beta_0 + \beta_1 ln(Vehicles) + \beta_2 ln(Higheds) \\ +\beta_3 Children + \beta_4 ln(Income) + \beta_5 ln(GRP) + \beta_6 NewParts2 + \epsilon \end{aligned} \tag{4}$$

The leverage from Model-2e against the linear predictor are shown the figure 2, from which the influence of observations are on the predicted values could be shown. Two reference lines are put at $y = 1/n$ and $y = 2(p+1)/n$, where n is the number of observations and $(p+1)$ is the total number of the beta values or the predictor number plus one.

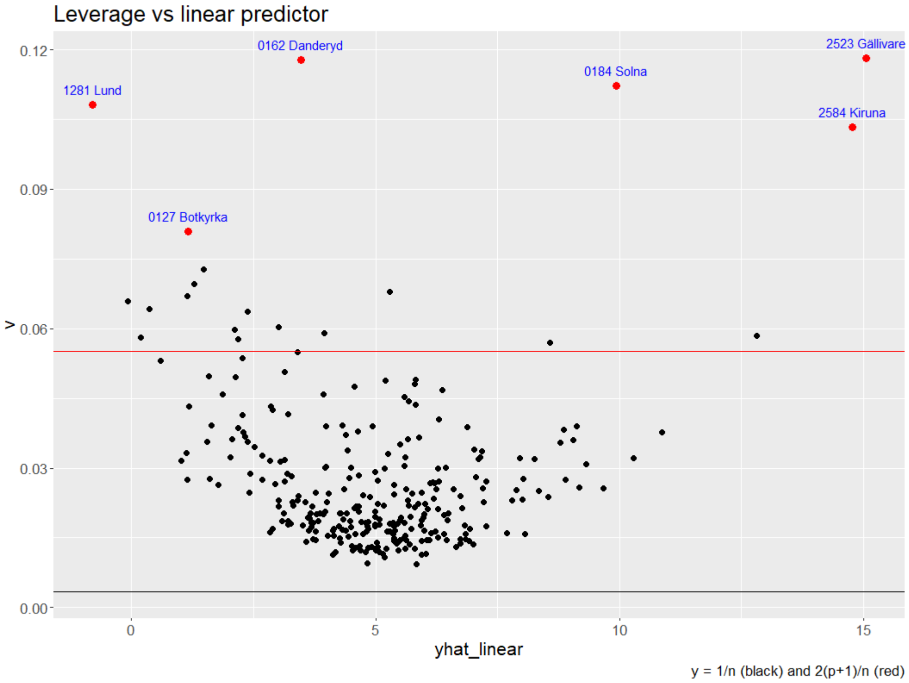

Figure 2: Leverage vs. Log predictor

From the figure 2, we could conclude that the six municipalities with the highest leverage are: 1281 Lund,0162 Danderyd, 0184 Solna, 2523 Gaellivare, 2584 Kiruna, 0127 Botkyrka. 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),NewParts2.

The pairs of x-variables against each other and separately for each category are plotted below, separated by NewParts2.The six municipalities with highest leverage are highlightened.

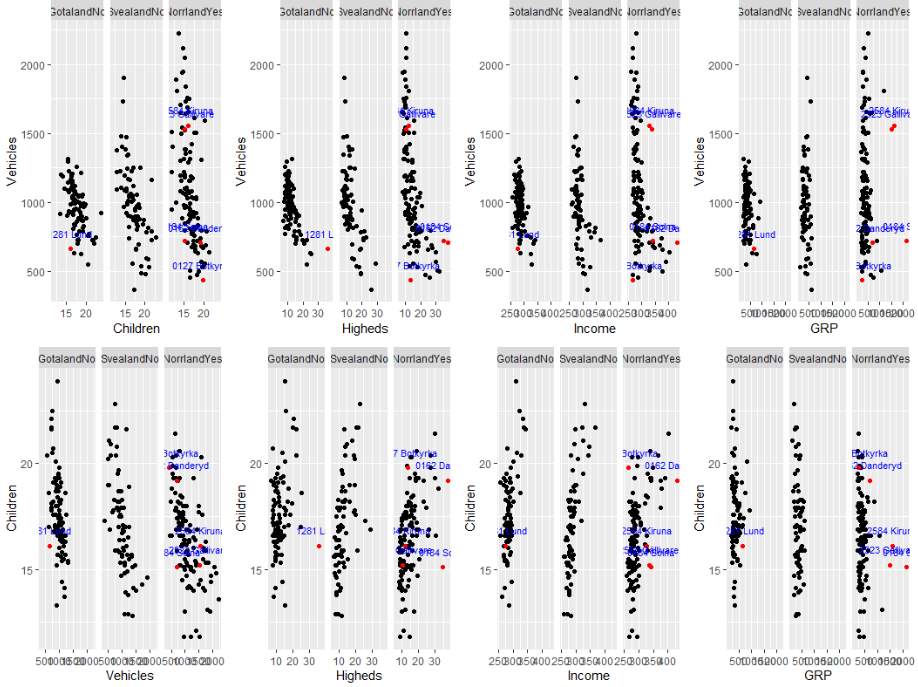

Figure 3: Vehicles,Children against other variables

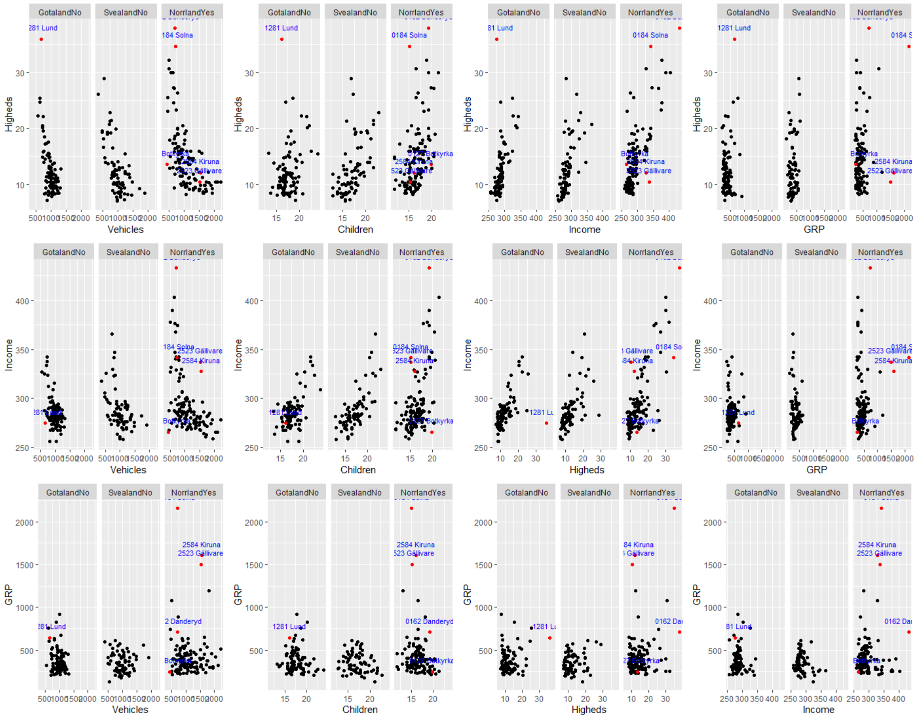

Figure 4: Higheds,Income,GRP against other variables

From figures above,it can be seen that the GRP, Income, and Higheds remain stay far away from the x-variables' center of gravity, which suggests they contributed to the leverage more than other variables,and thus have huge influence on the prediction.

3.(d) Explain,exclude,refit

In this section, the following steps shall be executed repeatedly until no municipalities with large residual with identifiable emission sources are left.

1 Identify the municipalities where $r_i^* > +3$ and try to identify the emission source.

2 Remove these identified municipalities from the data and refit the model.

First, the observations with studentized residual r_i^* larger than 3 and have identifiable emission sources are:

Municipality with $r_i^* > 3$	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

Table 2: Removed Municipalities-Round 1

In the second round, all the data listed above have been removed, the new model was updated yet high-residual municipalities still emerged on the square-root studentized residual plot. Among these high-residual municipalities, the ones with identifiable sources are listed in table 3.

Municipality with $r_i^* > 3$	Emission Source or Company name
1882 Askersund	Ahlstrom
1484 Lysekil	Preeem AB
1761 Hammarö	paper mill
1480 Göteborg	Preem AB,Stl Refinery AB
1494 Lidköping	Swedish Air Force wing
2284 Örnsköldsvik	Metsa Board Sverige AB

Table 3: Removed Municipalities-Round 2

In the third round, we repeated the procedures, except for one municipality ”1471 Götene” don't have identifiable source, the rest of the high-residual municipalities with emission source are as follows:

Municipality with $r_i^* > 3$	Emission Source or Company name
0319 Alvkarleby	Stora Enso
2262 Timrå	paper mill
1781 Kristinehamn	Nordic paper
1460 Bengtsfors	Ahlstrom

Table 4: Removed Municipalities-Round 3

In the fourth round, another 3 municipalities are about to be removed.

Municipality with $r_i^* > 3$	Emission Source or Company name
0980 Gotland	Cementa AB/Heidelberg Materials
1272 Bromölla	Stora Enso
1885 Lindesberg	Billerud AB

Table 5: Removed Municipalities-Round 4

After four rounds of exclusions on the municipalities, except for one municipality "1471 Götene", there exists no more ambient data point in the square root of the studentize-residual plot. All the removed municipalities are listed below in table 6:

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
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 6: 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 7 and table 8.

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 7: 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 8: 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.

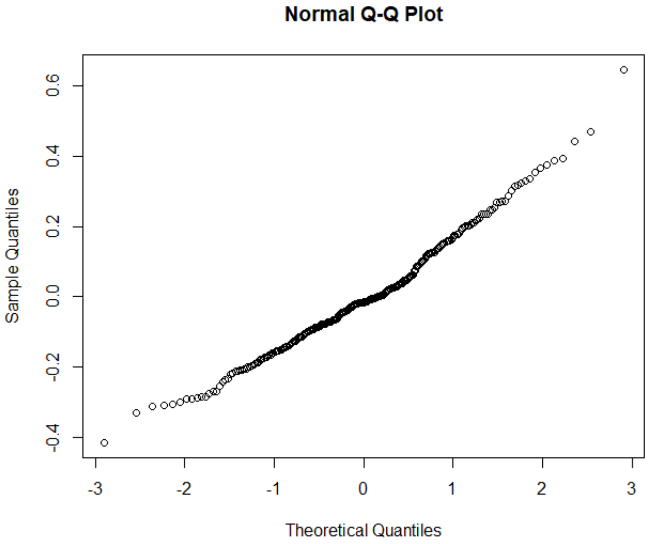

Figure 5: Q-Q plot of the new model

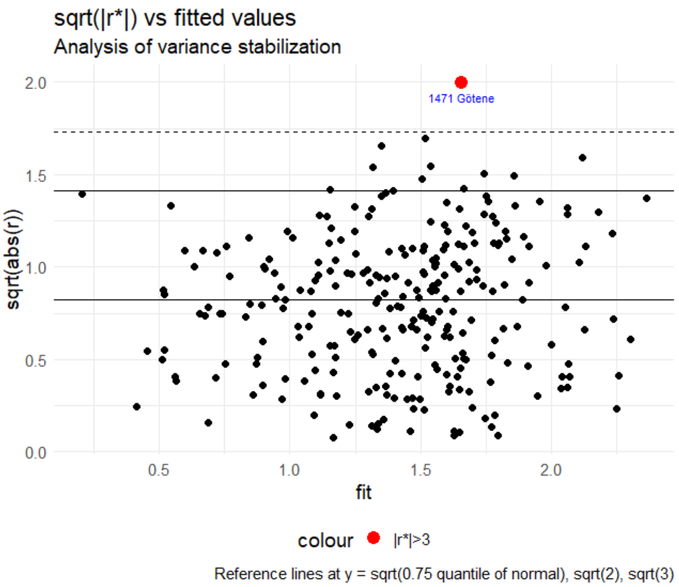

Figure 6: Square root of the absolute studentized residuals

3.(e) Variable selection

In this section, the clean data are used to select the best model, 4 models would refit with the cleaned data: model-1b,model-2c,model-3d, and the null model.To perform the stepwise selection using AIC or BIC as a criterion from both directions, we could set the starting point to be model-1b, the lower scope to be model-null, the upper scope to be model-3d, the parameter trace to be TRUE.

AIC criterion: In terms of AIC stepwise selection, at each step, both backward elimination and forward selection would be conducted and the action that can cause the largest decrease in AIC would be selected and become the base model of the next step.

Each step along with the choice of the variables will be fully discussed below.

Step1:Forward Selection. Base the model-1b($log(PM10) \sim log(Vehicles)$),the AIC printed out for model-1b is -885.46, shown below as operation ”none” . the AIC values of all backward elimination and forward selection are listed below(Note the AIC values displayed by setting trace=”TRUE” differ from using AIC(). However, in general it still conveys effective information in variable selection procedure): It can be seen that adding log(Income) could minimize the AIC value(-926.66), so we use $log(PM10) \sim log(Vehicles) + log(Income)$ as the base model in the next step.

	Df	Sum of Sq	RSS	AIC
+ log(Income)	1	1.499	8.676	-926.66
+ log(Higheds)	1	1.241	8.934	-918.72
+ NewParts	2	0.831	9.344	-904.55
+ Children	1	0.730	9.446	-903.63
<none>			10.176	-885.46
+ log(GRP)	1	0.005	10.170	-883.60
- log(Vehicles)	1	41.382	51.558	-447.70

Step2:Forward Selection. Base the model $log(PM10) \sim log(Vehicles) + log(Income)$,the operation which could minimize the AIC value is adding NewParts2,since after adding NewParts2, the AIC value could drop to -941.45. **Step3:Forward Selection.** Base the model $log(PM10) \sim log(Vehicles) + log(Income) + NewParts2$,the operation which could minimize the AIC value is adding Children,since after adding Children, the AIC value could drop to -954.62.
Step4:Forward Selection. Base the model $log(PM10) \sim log(Vehicles) + log(Income) + NewParts2 + Children$,the operation which could minimize the AIC value is adding log(Higheds),since after adding log(Higheds), the AIC value could drop to -962.

To conclude, when using AIC criterion conducting automatic stepwise selection online, the best model would be: $log(PM10) \sim log(Vehicles)+log(Income)+NewParts2 + Children + log(Higheds)$.The final AIC value of the model is -190.9.
BIC criterion: Base the model-1b($log(PM10) \sim log(Vehicles)$), the BIC printed out for model-1b is -878.25, shown below as operation "none". the BIC values of all backward elimination and forward selection are listed below(Note the BIC values displayed by setting trace="TRUE" differ from using BIC(). However, in general it still conveys effective information in variable selection procedure): It can be seen that adding log(Income) could minimize the BIC value(-915.85), so we use $log(PM10) \sim log(Vehicles) + log(Income)$ as the base model in the next step.

	Df	Sum of Sq	RSS	AIC
+ log(Income)	1	1.499	8.676	-915.85
+ log(Higheds)	1	1.241	8.934	-907.91
+ Children	1	0.730	9.446	-892.82
+ NewParts	2	0.831	9.344	-890.15
<none>			10.176	-878.25
+ log(GRP)	1	0.005	10.170	-872.79
- log(Vehicles)	1	41.382	51.558	-444.10

Step2:Forward Selection. Base the model $log(PM10) \sim log(Vehicles) + log(Income)$,the operation which could minimize the BIC value is adding NewParts2,since after adding NewParts2, the BIC value could drop to -923.44.

Step3:Forward Selection. Base the model $log(PM10) \sim log(Vehicles) + log(Income) + NewParts2$,the operation which could minimize the BIC value is adding Children,since after adding Children, the BIC value could drop to -933.01.
Step4:Forward Selection. Base the model $log(PM10) \sim log(Vehicles) + log(Income) + NewParts2 + Children$,the operation which could minimize the BIC value is adding log(Higheds),since after adding log(Higheds), the BIC value could drop to -936.79.
Step5:Backward Elimination. Base the model $log(PM10) \sim log(Vehicles)+ log(Income) + NewParts2 + Children + log(Higheds)$, the operation which could minimize the BIC value is eliminating log(Income), since after this, the BIC value could drop to -937.68. To conclude, when using BIC criterion conducting automatic stepwise selection online, the best model would be: $log(PM10) \sim log(Vehicles) + NewParts2 + Children + log(Higheds)$.The final BIC value of the model is -162.01.
Conclusion In the table below 9, all the measures of the models are listed for model selection.

Model	β	Residual Standard Deviation	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 9: Comparison between models

To assess how much the variability of log(PM10) can be explained,the adjusted R^2 could do. Higher adjusted R^2 stands for higher variability,so the best model should have the highest adjusted R^2. Besides, we also look for the lowest AIC and BIC value in the best model, since they show the tradeoff between small residual error and the large number of coefficients. To conclude, the AIC model is the best one selected by software. Additionally, it makes sense the best AIC model could best depict the PM10 emission. To be specific, people living in different regions with different climates may affect the PM10 level, for example, people in Lund favor biking everywhere they go because the city is small and bikes are cost-efficient for students who consist of a big portion of the population here. Besides, a higher income level or higher educational level could also more possibly drive people to live a more sustainable lifestyle thus reducing the PM10 emission. Furthermore, more children means fewer vehicle owners. Finally, it's obvious more vehicles would cause more emissions.