LR project2

Wanru Cheng, Kuan Teh Wan

May 2024

Introduction

Project 2 mainly focuses on modeling how the probability of a municipality having a high number of cars per capita varies as a function of one or several variables , $\mathbf{x_i} = (x_{i1}, ...x_{ip})$.

Part 1. Introduction to Logistic Regression

1.(a) High number of cars without regression

In this section, we are asked to the probabilities p, odds p/(1-p),log odds of having high number of cars for each of the three Parts: Götaland, Svealand, and Norrland. Besides, the odds ratios and log odds ratios for Svealand and Norrland against Götaland are also in request. All of the results are listed below in the table 1.

Table 1: Results for high number of cars without regression

	probability p	odds p/(1-p)	log odds	odds ratio(ref: Götaland)	log odds ratio
Götaland	0.114	0.129	-2.048	-	-
Svealand	0.208	0.263	-1.335	2.039	0.713
Norrland	0.407	0.688	-0.375	5.328	1.673

From the table listed above, the odds of having a high number of cars would increase drastically by 203.9%. Additionally, the plot in which the car category versus GRP is displayed in figure 1.

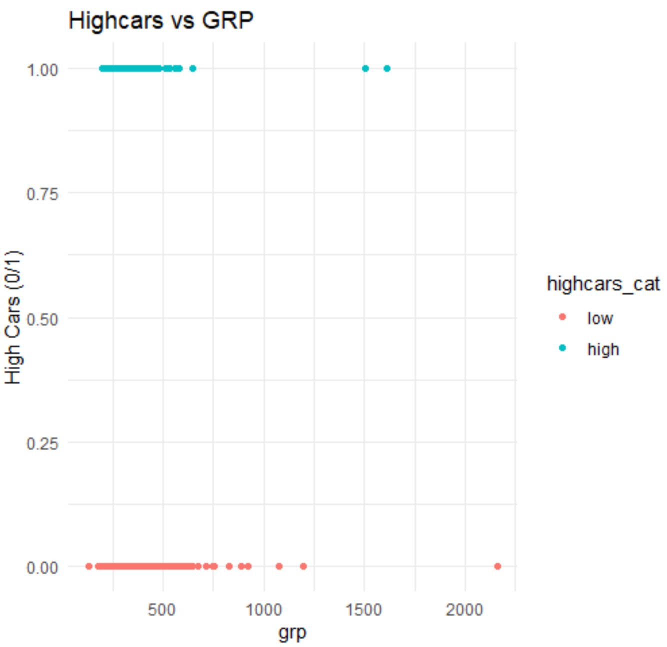

Figure 1: Highcars vs. GRP

1.(b) High number of cars with regression

We fitted a logistic model, *Model.1(b)* with Part as the explanatory variable:

$$ln\frac{p_i}{1-p_i} = \beta_0 + \beta_1 \text{Part} \tag{1}$$

Model.1(b)'s $\beta-$estimates, the standard errors, corresponding e^{β} and all confidence intervals are shown in the table below:

	Intercept	Sveland	Norrland
β	-2.05	0.71	1.67
SE	0.27	0.28	0.38
95% C.I.	(-2.61,-1.56)	(0.00,1.44)	(0.93,2.44)
e^{β}	0.13	2.04	5.33
SE	1.30	1.40	1.47
95% C.I.	(0.07,0.21)	(1.00,4.23)	(2.53,11.48)

Table 2: *Model.1(b)* $\beta-$estimates, standard errors and 95% confidence intervals

We notice that odds for Norrland and odds ratio for Sveland and Norrland are the same as what we have calculated in the table in 1(a).

Next, the estimations for the log-odds of high numbers of car, with the standard errors and confidence intervals are shown below.

	$\hat{p}$	logit.fit	logit.se.fit	logit.lwr	logit.upr
Gotaland	0.11	-2.05	0.27	-2.57	-1.53
Svealand	0.21	-1.34	0.25	-1.83	-0.84
Norrland	0.41	-0.37	0.28	-0.92	0.17

Table 3: Results for the log-odds of high numbers of car

	odds.lwr	odds.upr	p.lwr	p.upr
Gotaland	0.08	0.22	0.07	0.18
Svealand	0.16	0.43	0.14	0.30
Norrland	0.40	1.18	0.29	0.54

Table 4: Results for the log-odds of high numbers of car

At last, we want to test whether there are any significant differences in the probability of a large number of cars between the three parts of Sweden. For that we are using the LR-test against the null model. We have the hypothesis H0 : $\beta_{Part} = 0$. For this we take the difference of the deviance of the null model and our Model.1(b) and compare it to the χ^2-quantile.

$$D_0 - D = 19.47479 > \chi^2_{0.05}(2) = 5.991465$$

Now, we can say with at least a 95% probability that the variable Part is significant to describe the amount of highcars.

1.(c) Access to a bus stop

To start with, we plotted 0/1 variable highcars against Transit to observe whether the proportion of the inhabitants that live less than 400m to bus stops would by itself explanatory enough on inhabitants' high number of cars per capita.

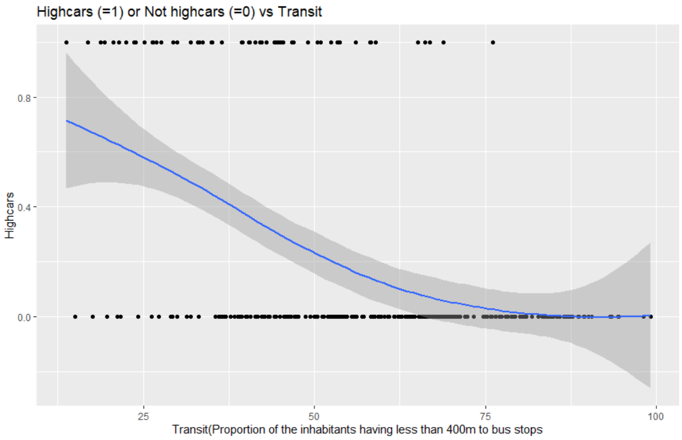

Figure 2: moving average of highcars against Transit

Following that, we fitted Model.1(c) using Transit as an explanatory variable. The β-estimates and e^{β}-estimates along with their 95% confidence intervals are reported below.

Table 5: β-estimate and e^{β}-estimate along with the confidence intervals

	β	ci.lower-2.5%	ci.upper-97.5%	e^{β}	ci.lower-2.5%	ci.upper-97.5%
(Intercept)	2.38	1.39	3.46	10.83	4.01	31.67
Transit	-0.07	-0.10	-0.05	0.93	0.91	0.95

After that,we added the estimated probability of a high number of cars and its confidence level to the former plot in figure 3.

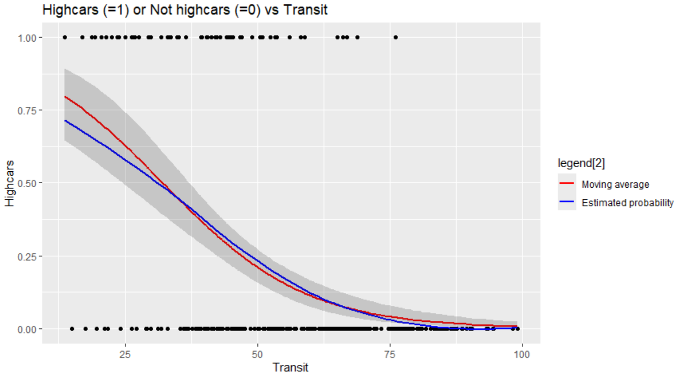

Figure 3: estimated probability and moving average of highcars against Transit

Kommuner dataset only consists of 290 samples, so it would be problematic to conduct Wald test on it. The likelihood ratio test is utilized here to gain the more precise result. According to the definition of the global likelihood ratio test, if H_0 is true, asymptotically as n $\to \infty$, $D_0 - D \sim \chi^2_\alpha(p)$, and we should reject H_0 at significance level $\alpha(0.95)$, if $D_0 - D > \chi^2_\alpha(p)$. In this case, the null hypothesis H_0 is $\beta_1 = 0$ and the degree of freedom is 1,thus $D_0 - D \sim \chi^2_\alpha(1)$ when H_0 is true and asymptotically as n $\to \infty$.The test statistic $D-diff$, $chi2_\alpha$ and $P - value$ are listed in the table 6.

Table 6: Test statistics of LR test

D-diff	Chi2-alpha	P-value
67.74	3.84	1.87e-16

From the table 6, it is clear that the difference between D_0(deviance of null model) and D(deviance of this model) is much bigger than the χ score at significance level 0.95. Furthermore, the P-value is much smaller than the χ score at a significance level 0.95, which suggests that the current model provides a significantly better fit than the null model.

Finally, if the proportion living close to a bus stop changes,the odds of of having a high number of cars would change as what showcased in table 7. pu is short for percentage unit. Given the value of β_1 from the model summary being -0.07417,the change of the odds could be calculated via $e^{\beta_1 * n}$,where n equals the change of the the percentage unit of x-variable.

Table 7: how highcars change as x-variable changes

+1 pu	+10 pu	-1 pu	-10 pu
-7.15%	-52.37%	7.70%	109.95%

1.(d) Leverage

Here, we have calculated the leverage values for Model.1(c) and plot them against Transit. Figure as shown below.

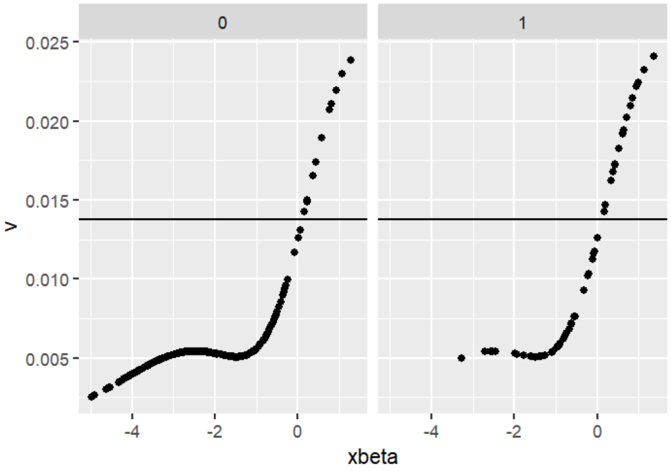

Figure 4: Plot of Model.1(c) leverage against Transit

By observing the graph, we can see two notable bumps. These occur where the S-curve changes its slope, indicating regions where individual data points exert significant influence on the estimation of regression coefficients. The location of these ”bumps” in leverage corresponds to regions where the relationship between Transit and the likelihood of high car ownership transitions from steep to shallow or vice versa.

1.(e) Calculate McFadden’s adjusted pseudo R^2,AIC,BIC for Model.1(b) and Model.1(c)

Recalled that the bigger McFadden’s adjusted pseudo R^2 is for a model, the better the model fit is, what’smore, the model with smallest AIC and BIC value is what we decide to be the best. McFadden’s adjusted pseudo R^2,AIC, and BIC for Model.1(b) and Model.1(c) are shown in the table 8.

Table 8: Model comparison statistics

	R^2_{MF}	R^2_{MFadj}	AIC	BIC
Model.1(b)	0.00	-0.0037	274.91	282.25
Model.1(c)	0.1787	0.1750	226.50	233.84

From the table above, it could be concluded that Model.1(c) is a better fit,which is access to a bus stop is more explanatory of the probability of having a high number of cars in a municipality.

Table 9: The VIF values of Model.full

Children	6.343
Seniors	5.077
log(Higheds)	1.992
log(Income)	2.791
log(GRP)	1.613
Urban	2.233
Transit	1.850
log(Apartments)	3.058
Fertility	1.581
Persperhh	5.669

Part 2. Variable selection and influential observations

In section 2, we will polish the data further to select the most explanatory variables for the model, such as imputation of NA value in Fertility, AIC and BIC stepwise selection

2.(a) Imputation of missing data

Since we may want to use the Fertility rate in our model, we need to deal wit the missing values. To impute the NA value in the Fertility column, we could replace all the NA values by the mean value of Fertility, which is 1.57.

2.(b) Variable selection

After fitting a full model named Model.full with suitable log-transformed variables, we utilized it as the upper bound in stepwise AIC and BIC selections. We log-transformed the variables: Income,GRP,Apartments, and Higheds. The VIF values of Model.full are reported in the table 9.

From the table 9, it is clear that many variables such as Seniors,Persperhh,Children have such a large VIF(larger than 5), which suggests that the the amount of variability that can be explained by other x-variables has reached over 80%. The Model.full is thus a bad fit and we should wisely exlude msome variables to improve the overall performance.To achieve this goal, we conducted AIC and BIC stepwise selections and variables included in their final versions Model.AIC and Model.BIC and their corresponding VIF-values are displayed below in table 10.

From the table above, we could now say that the multicolinearity problems have been greatly relieved, most of the VIF values in these two models are around 2.0, which suggests their variabilities can be expressed by other variables by around 50%. Following that, we conducted partial LR-test on variables Fertility and Transit. The null hypothesis H_0 is: the β values of Fertility and

Table 10: VIF-values of AIC and BIC models

	Urban	log(Apartments)	Persperhh	log(Income)	Fertility	Transit
Model.AIC	2.157	2.080	2.294	2.059	1.282	1.650
Model.BIC	1.735	1.934	2.186	2.008	-	-

Transit equal 0,and if H_0 is true,then asymptotically $D_bic - D_aic \sim \chi^2_\alpha(2)$ fulfills. The test statistics are displayed as follows.

Table 11: Test statistics on Model.AIC and Model.BIC

D-diff	Chi-alpha	P-value
6.150	5.991	0.046

From the statistics above, it can be concluded that $D_bic - D_aic > \chi^2_\alpha(2)$,thus we should reject H_0.Furthermore, P-value is smaller than 0.05 which suggests that it would be better off with the additional variables Transit and Fertility. Finally, we examined the McFaden's adjusted R^2,AIC and BIC on both models, which are all listed below in table 12.

Table 12: Model AIC and BIC comparison statistics

	R^2_{MF}	R^2_{MFadj}	AIC	BIC
Model.AIC	0.461	0.440	170.38	196.07
Model.BIC	0.440	0.426	172.53	190.88

From the statistics above, it could be seen that Model.AIC is better fit since it has larger adjusted McFaden's R^2 and smaller AIC value. Thus Model.2(b) should take Model.AIC.

2.(c) Influential observations

In this subsection, we investigated the influence of observations on the fitted values.In the figure 5, we plotted the leverage for Model.2b against the linear predictor, with the reference line located at $y = 2(p+1)/n$.

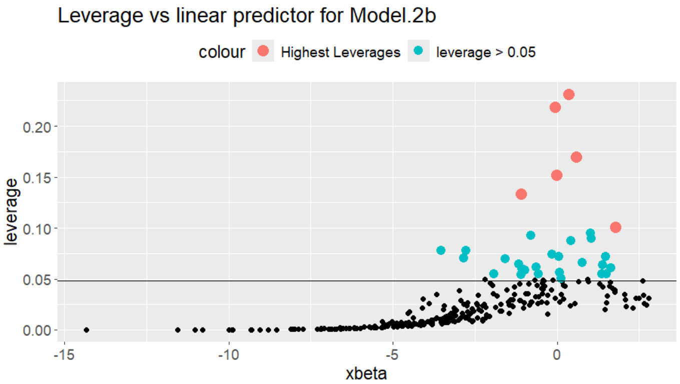

Figure 5: leverage against linear predictor

From the figure 5, we have highlighted the top 6 municipalities in red and all municipalities which is bigger than $y = 2(p+1)/n$. The municipalities with a high leverage are: 2523 Gällivare, 2422 Sorsele, 1419 Tjörn, 2418 Mala, , 0512 Ydre, 2026 Gagnef. After that, we plotted the cookäs distance with the reference line at both $2/\sqrt(n)$ and 1.

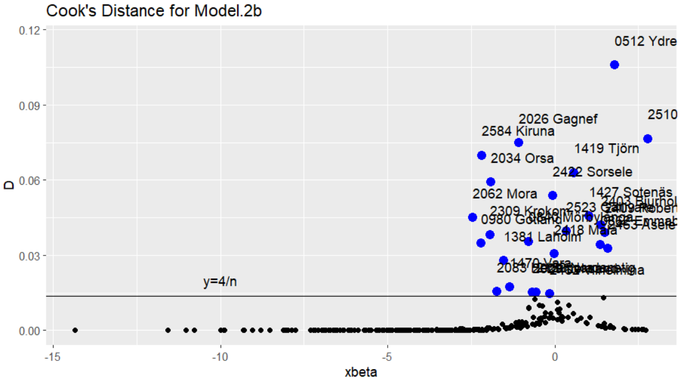

Figure 6: Cook's distance for Model.2b

From figure 6, it can be seen that the 6municipalities with highest cook's distance are: 0512 Ydre, 0840 Moerylanga, 0862 Emmaboda,0980 Gotaland,1381 Laholm, 1419 Tjörn. Finally, we investigated the DFBETAS and plotted the figures with DFBETAS of each β parameter against fitted values. The red points are those with highest cook's distances. It could be concluded that for almost all the parameters, municipalities with high cook's distances have a large influence on β parameters.

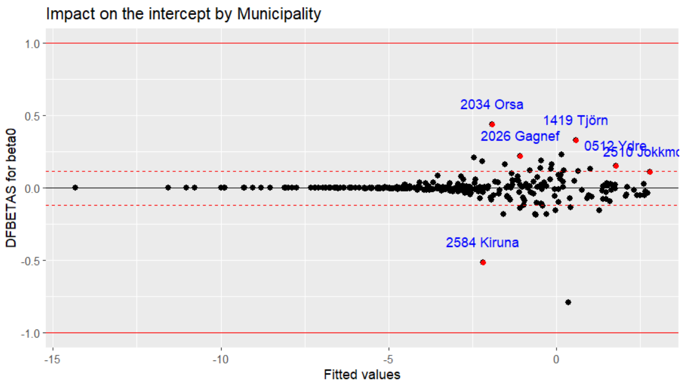

Figure 7: Influence on Intercept

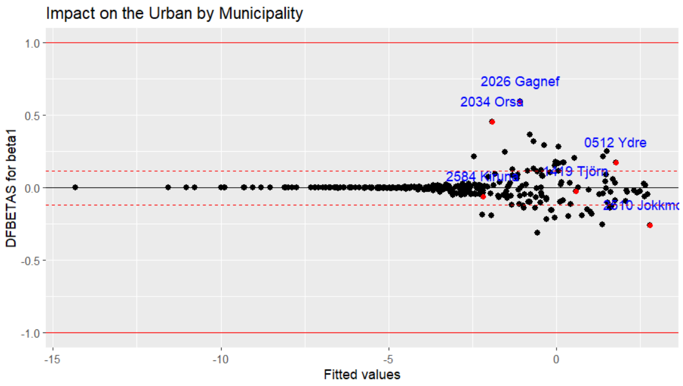

Figure 8: Influence on $\beta 1$

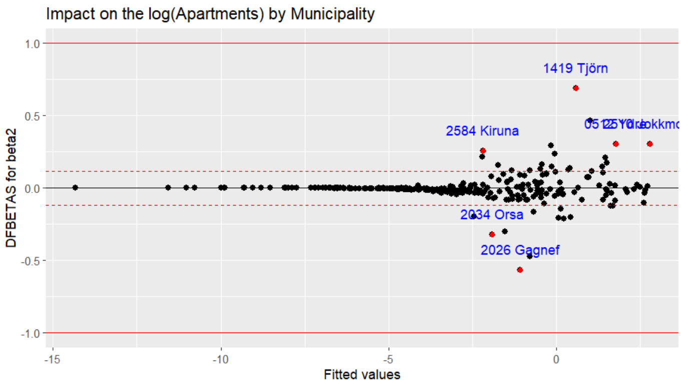

Figure 9: Influence on $\beta 2$

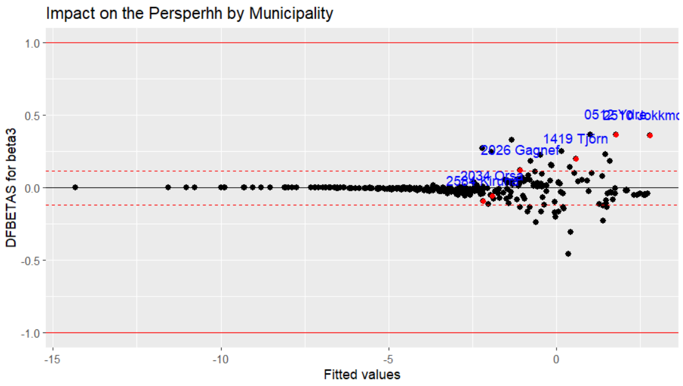

Figure 10: Influence on $\beta 3$

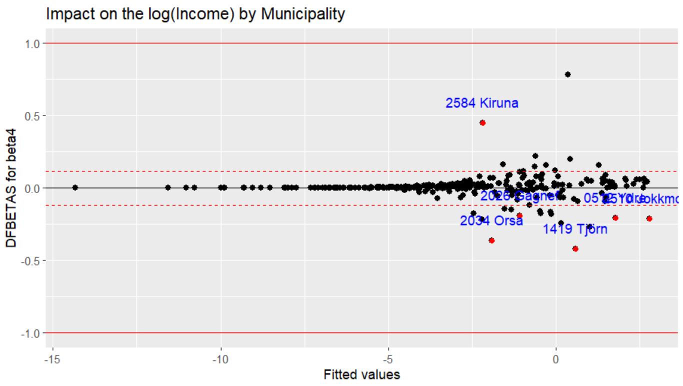

Figure 11: Influence on $\beta 4$

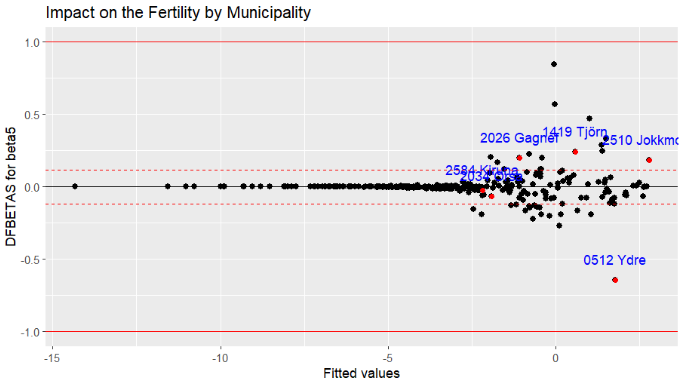

Figure 12: Influence on $\beta 5$

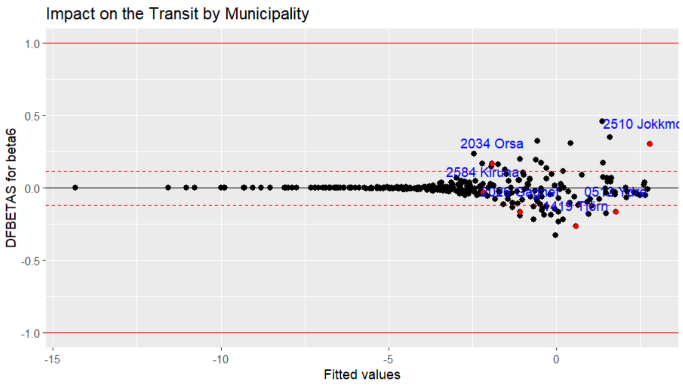

Figure 13: Influence on $\beta 6$

2.(d) Deviance residuals

In this subsection, we plotted the standardized deviance residuals against the linear predictor. The reference lines are located in $y = 2, -2, 3, -3$. We highlighted the top 6 observations with high Cook's distances in red. Since there exists no observations with large deviance residuals, so none was shown on the plot.

It could be seen that the municipalities with high cook's distance have a larger deviance residuals whose absolutes are around 2. This shows the high influence of the problematic municipalilties.

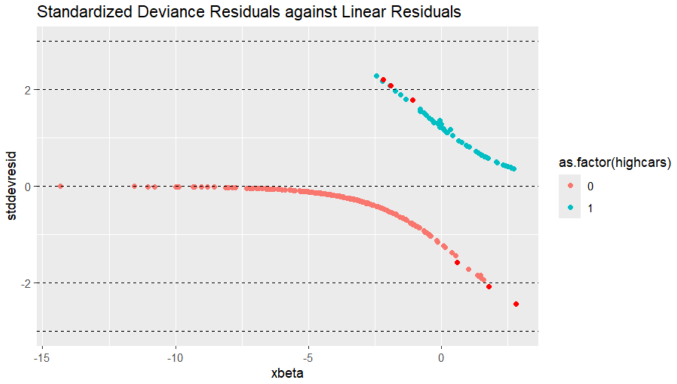

Figure 14: Standardized Deviance Residuals against the linear predictor

Finally, we plotted the standardized deviance residuals against each x-variables.

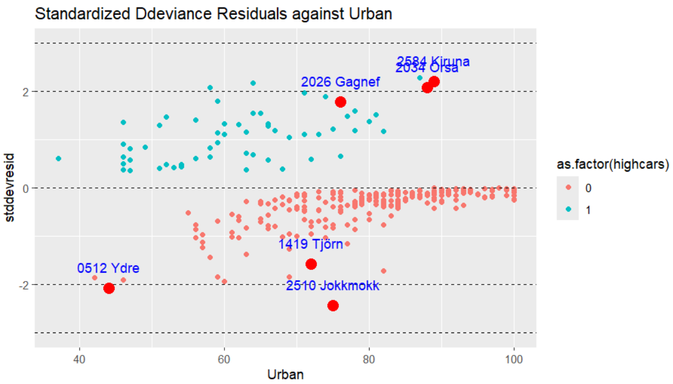

Figure 15: standardized deviance residuals against Urban

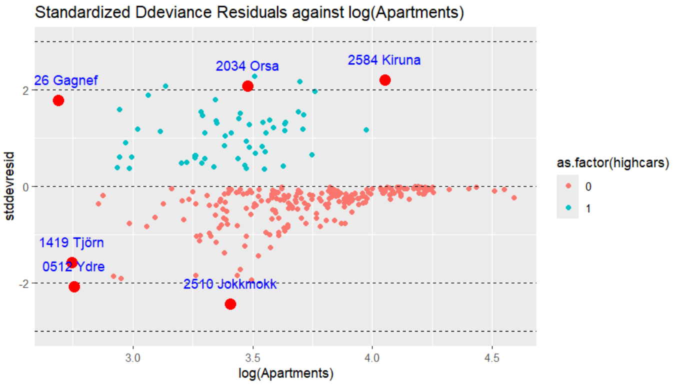

Figure 16: standardized deviance residuals against log(Apartments)

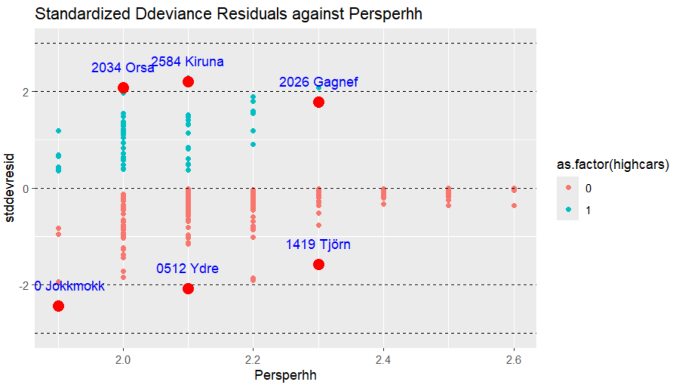

Figure 17: standardized deviance residuals against Persperhh

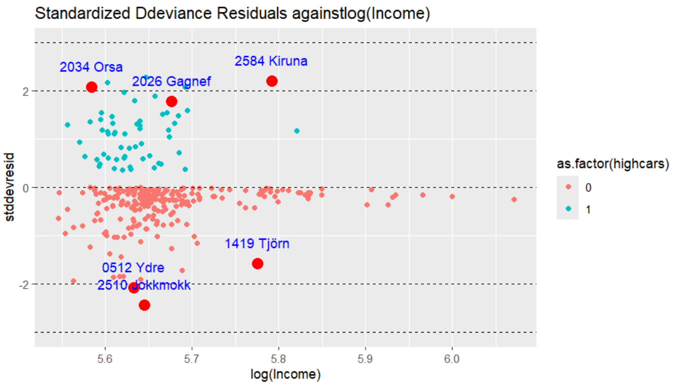

Figure 18: standardized deviance residuals against log(Income)

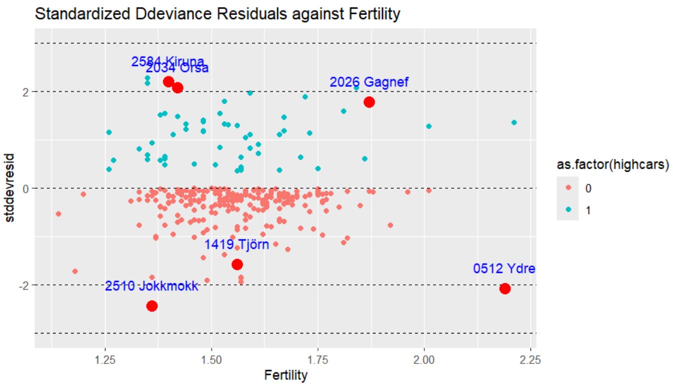

Figure 19: standardized deviance residuals against Fertility

From figure 15 to figure 19, we could see that all municipalities with high cook's distance lie far away from the gravity of x-variables. Additionally, every plot has a certain pattern and the standardized deviance residuals are not randomly distributed around 0. For example, for Urban,Persperhh,log(income), log(Apartments), the high car ownership gets larger as x-variables get larger, while the low car ownership approaches 0 as x-variables get larger.

1 Part 3 Goodness-of-fit

3.(a) Confusion

Confusion Matrices for all models:

	Actual:Low	Actual:High
Predicted: Low	232	58
Predicted: High	0	0

Table 13: Confusion Matrix for *Model.null*

	Actual:Low	Actual:High
Predicted: Low	232	58
Predicted: High	0	0

Table 14: Confusion Matrix for *Model1(b)*

	Actual:Low	Actual:High
Predicted: Low	219	40
Predicted: High	13	18

Table 15: Confusion Matrix for *Model1(c)*

	Actual:Low	Actual:High
Predicted: Low	217	25
Predicted: High	15	33

Table 16: Confusion Matrix for *Model.AIC*

	Actual:Low	Actual:High
Predicted: Low	215	26
Predicted: High	17	32

Table 17: Confusion Matrix for *Model.BIC*

	Actual:Low	Actual:High
Predicted: Low	226	7
Predicted: High	6	51

Table 18: Confusion Matrix for *Model.full*

The Goodness-of-fit statistics are presented below for all 6 models

Model	Accuracy	P Value Acc	Cohen Kappa	P Value McNemars	Sensitivity	Specificity
Null	0.8	5.35 e-01	0.00	7.18e-14	0.00	1.00
1b	0.8	5.35 e-01	0.00	7.18e-14	0.00	1.00
1c	0.82	2.57 e-01	0.31	3.55e-04	0.31	0.94

Model	Accuracy	P Value Acc	Cohen Kappa	P Value McNemars	Sensitivity	Specificity
BIC	0.85	1.43 e-02	0.51	0.222	0.55	0.93
AIC	0.86	3.83 e-03	0.54	0.154	0.57	0.94
Full	0.96	1.80 e-14	0.86	1	0.88	0.97

Table 19: Table3(a)

From the tables, we can conclude that Model.full, Model.AIC and Model.BIC are significantly better than those models always predicting the number of cars will be low(such as Model.null and Model 1(b) as their Specificity, true negative rate is 1) as they have a balanced value above 0.5 for both Sensitivity and

Specificity.

And among those models, Model.full is almost predicting the correct proportion of low and high number of cars as it has an accuracy of 0.96, close to 1.

3.(b) ROC-curves and AUC

Plot of all six models and a table for their AUC values and 95% confidence interval are presented below.

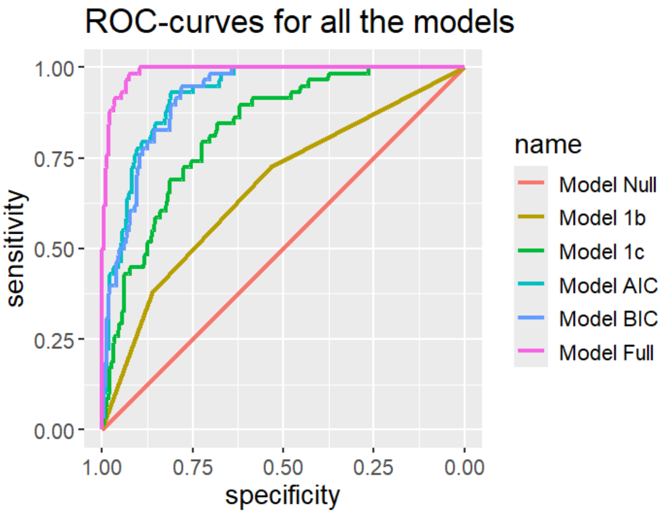

Figure 20: ROC Curves for all 6 models

Model	AUC	95% C.I.
Null	0.5000000	(0.5000000, 0.5000000)
1b	0.6676576	(0.5920988, 0.7432163)
1c	0.8274004	(0.7736394, 0.8811614)
AIC	0.9251635	(0.8945455, 0.9557815)
BIC	0.9203329	(0.8892679, 0.9513980)
Full	0.9886296	(0.9800940, 0.9971652)

Table 20: AUC values and their 95% confidence intervals

After performing the pair-wise tests, it does agree with the conclusion that we have, using AIC model as Model.2(b). As it has the closest value of test-statistic to zero against BIC model when comparing to all the other models.

3.(c) Optimal thresholds

After finding the optimal threshold, we come up with new confusion matrices for all models except for *model.null* and a new table with the following statistics:

	Actual:Low	Actual:High
Predicted: Low	124	16
Predicted: High	108	42

Table 21: Confusion Matrix for *Model1(b)*

	Actual:Low	Actual:High
Predicted: Low	194	9
Predicted: High	38	49

Table 22: Confusion Matrix for *Model1(c)*

	Actual:Low	Actual:High
Predicted: Low	189	9
Predicted: High	43	49

Table 23: Confusion Matrix for *Model.AIC*

	Actual:Low	Actual:High
Predicted: Low	220	5
Predicted: High	12	53

Table 24: Confusion Matrix for *Model.BIC*

	Actual:Low	Actual:High
Predicted: Low	226	7
Predicted: High	6	51

Table 25: Confusion Matrix for *Model.full*

Model	Accuracy	P Value Acc	Cohen Kappa	P Value McNemars	Sensitivity	Specificity
Null	0.8	5.35 e-01	0.00	7.18e-14	0.00	1.00
1b	0.57	1.00 e+00	0.16	3.03e-16	0.72	0.53
1c	0.74	9.96 e-01	0.39	4.91e-09	0.79	0.72

Model	Accuracy	P Value Acc	Cohen Kappa	P Value McNemars	Sensitivity	Specificity
BIC	0.82	2.11 e-01	0.54	4.73e-06	0.55	0.81
AIC	0.84	5.88 e-02	0.57	4.42e-05	0.57	0.84
Full	0.94	7.66 e-12	0.82	0.146	0.88	0.95

Table 26: Table3(c)

To discuss further, let's exclude the null and full model first. Although the accuracy has decreased for all the models we can still see improvements. The Cohen's kappa coefficient has increased for all the models, a good sign, which means that we are closer to 100% accuracy compared to random assignments.

The sensitivity and the specificity improved which comes from the minimization process. The P-value for the accuracy against the no information rate also got worse for all the models except for Model AIC. Thus, this confirms with our decision.

3.(d) Decision

In conclusion, we chose the best model for this problem will be the AIC model. It performed the best in the AIC-test, the Fadden's R2 adj, all the test on the confusion matrix and the area under the curve.

Here's the model's β-estimates and confidence intervals.

	β	95% C.I.
Intercept	-59.82	(-110.48,-7.78)
Urban	-0.05	(-0.10,0.00)
log(Apartments)	-0.11	(-0.18,-0.05)
Persperhh	-14.52	(-20.32,-9.48)
log(Income)	17.30	(6.96,27.94)
Transit	-0.03	(-0.06,0.01)

Table 27: Caption