LINEAR AND LOGISTIC REGRESSION

ALTERNATIVE A.
POISSON AND NEGATIVE BINOMIAL REGRESSION

YIFEI ZHANG

JIALU XU

Introduction

Our task is to fit a Poisson/Negbin model to represent the relationship between predictors and Cars_nbr. We will present in terms of the following points:

- Explore Explanatory Variables
- Data Transformation and Variable Selection
- Model Fitting and Evaluation
- Comparison and Model Selection
- Final Model Interpretation

LUND UNIVERSITY

Explanatory Variables

Categorical Variables (4st) : Kommun, County, Part, Coastal

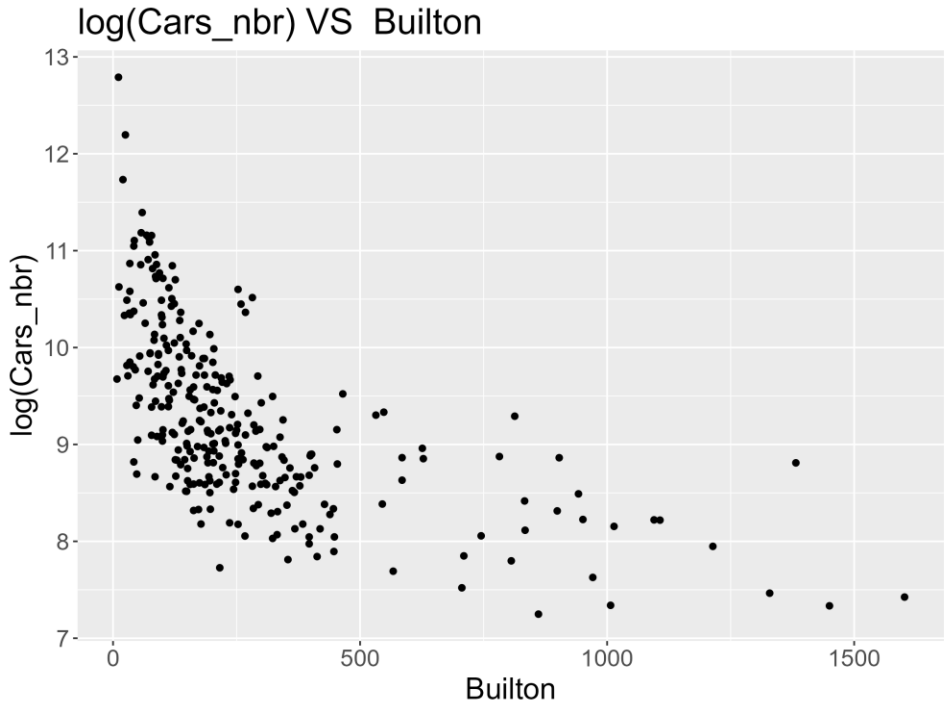
Numerical Variables (13st) : Vehicles, Builton, Children, Seniors, Higheds, Income, GRP, Urban, Transit, Apartment, Fertility, Persperhh, Population

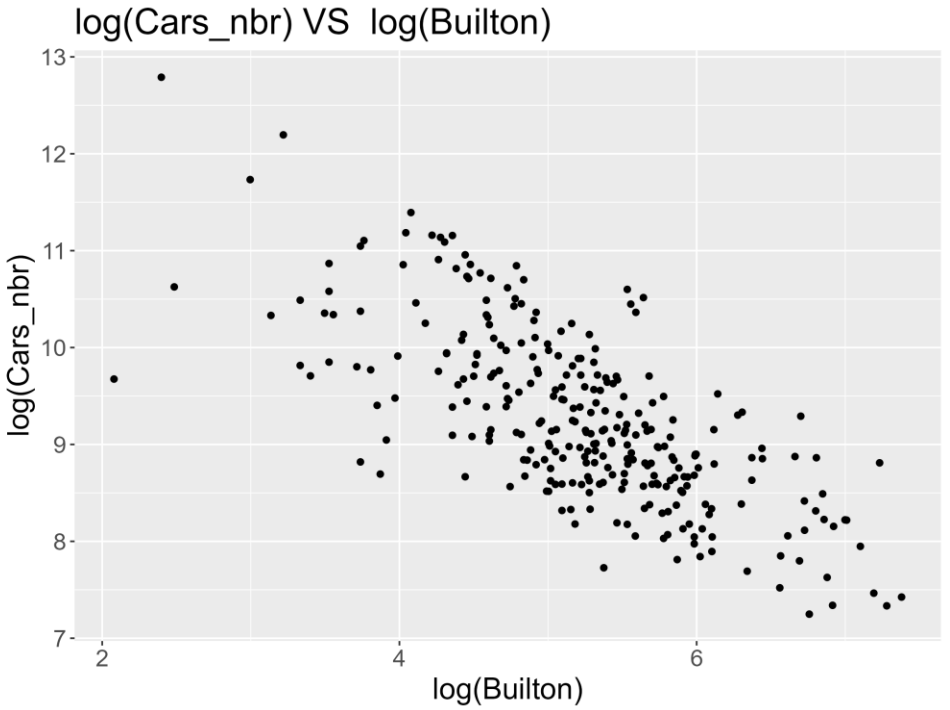
LUND UNIVERSITY

Log or not ?

We plot a scatterplot of log(Cars_nbr) vs Var / log(Var) to determine if we need to log transform the numerical variable.

Here we show some examples of plots:



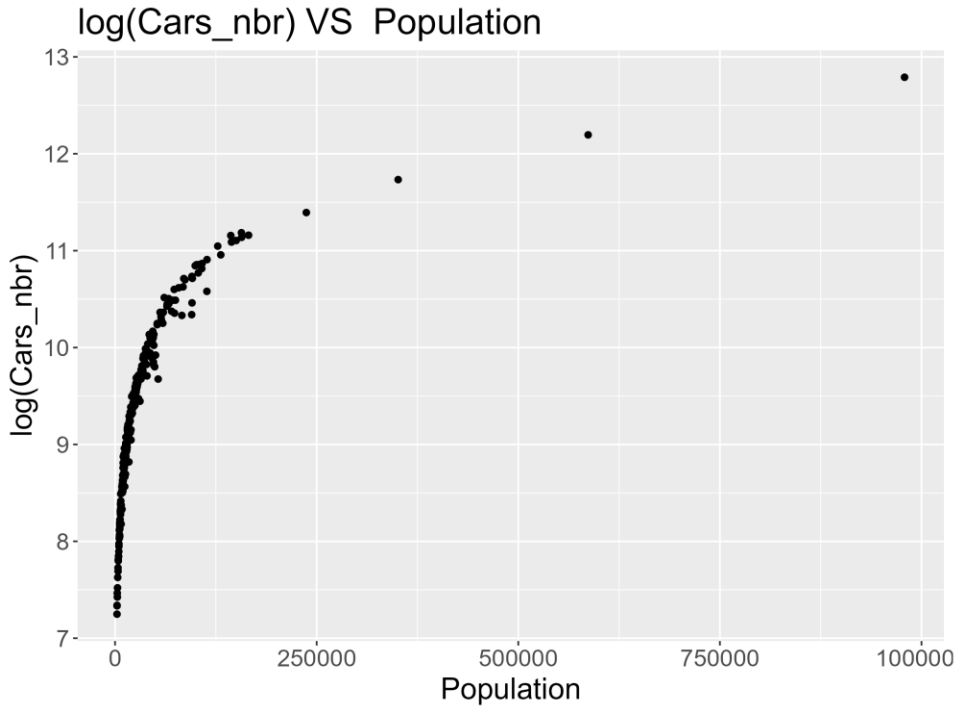


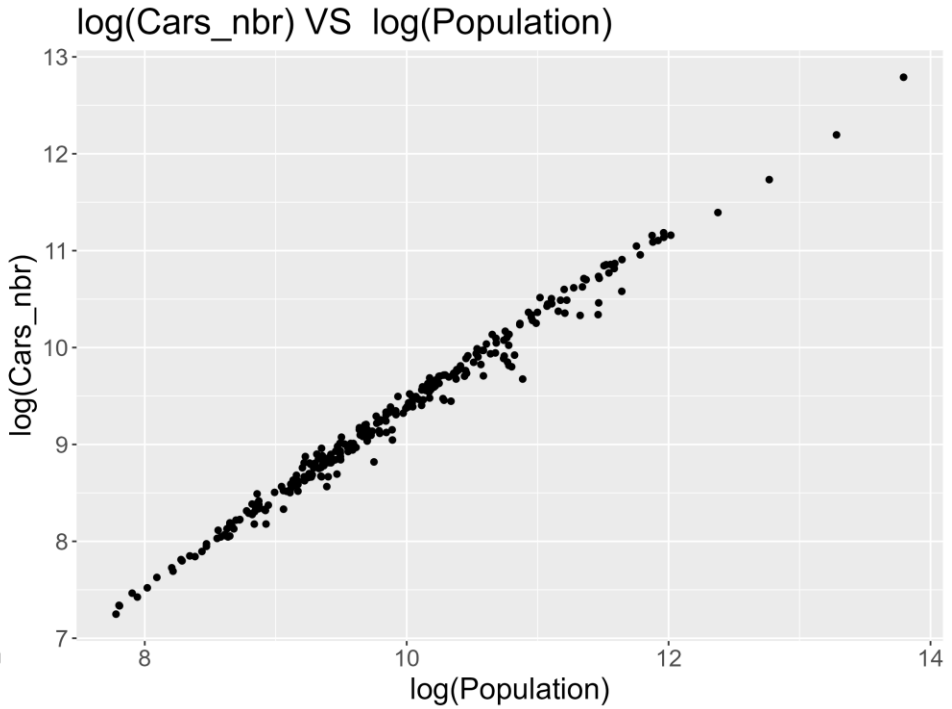
LUND UNIVERSITY

Log or not ?

We plot a scatterplot of log(Cars_nbr) vs Var / log(Var) to determine if we need to log transform the numerical variable.

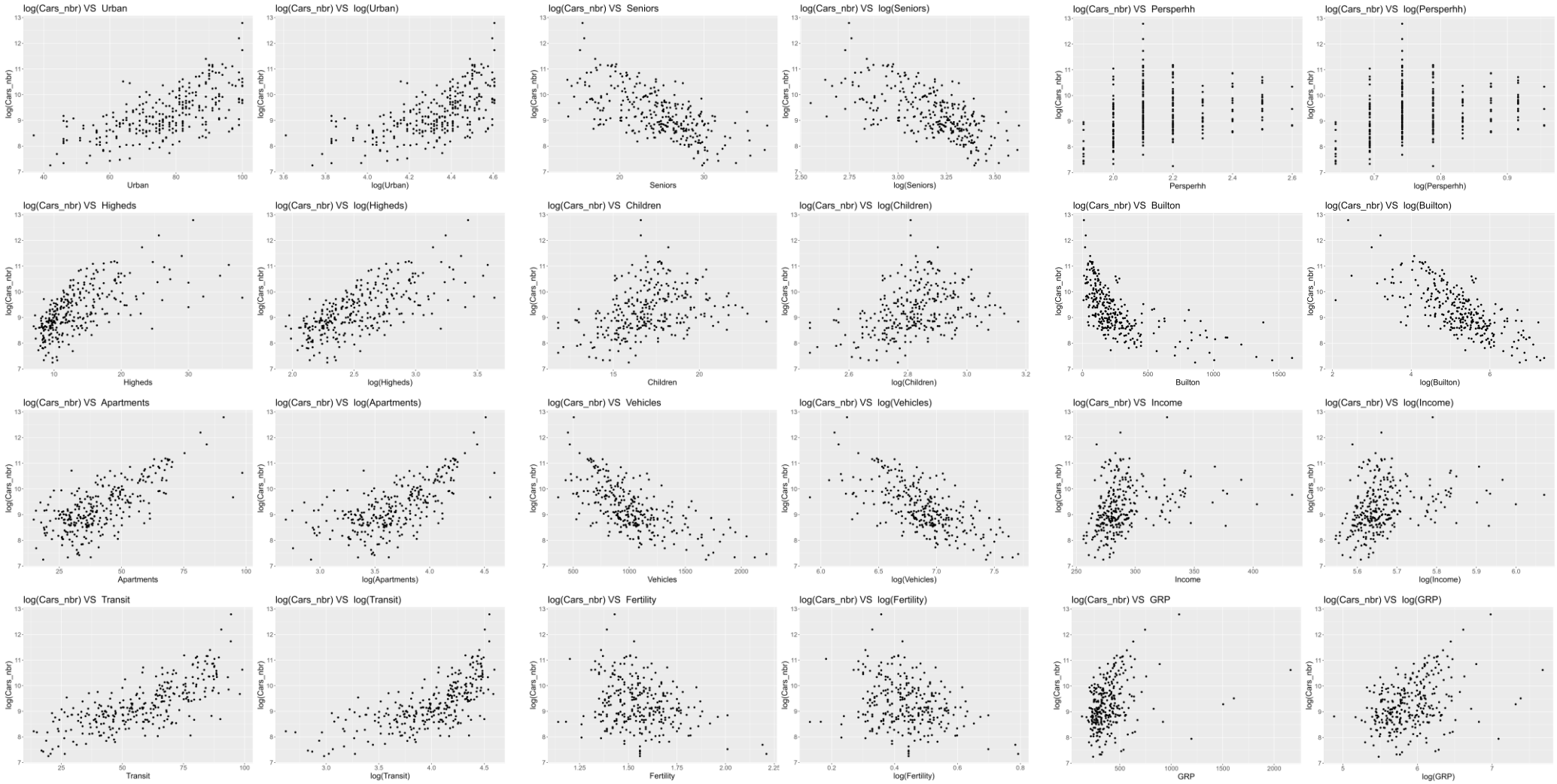
Here we show some examples of plots:





LUND UNIVERSITY

Log or not ?



LUND UNIVERSITY

Variable Selection

We selected all (log) numerical variables from Project1 & 2 as explantory variables by visualizing the scatterplot and intuition.

- log(Apartments)
- log(Builton)
- Log(Higheds)
- Log(Seniors)
- Transit
- Urban
- log(Vehicles)
- log(Children)
- log(Income)
- log(GRP)
- Fertility
- Persperhh

LUND UNIVERSITY

Model fitting – First Full Model

We first fit a Poisson regression model with all these variables.

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	18.55	0.05293	350.34	<2e-16
log(Apartments)	0.5176	0.003169	163.32	<2e-16
log(Builton)	-0.1495	0.001723	-86.81	<2e-16
log(Higheds)	1.171	0.002313	506.06	<2e-16
log(Seniors)	0.5255	0.006147	85.49	<2e-16
Transit	0.00927	5.209e-05	177.97	<2e-16
Urban	-0.005269	8.657e-05	-60.87	<2e-16
log(Vehicles)	-0.5206	0.004423	-117.69	<2e-16
log(Children)	0.7358	0.01129	65.17	<2e-16
log(Income)	-3.191	0.008913	-358.07	<2e-16
log(GRP)	0.5991	0.001708	350.68	<2e-16
Fertility	0.1387	0.005096	27.22	<2e-16
Persperhh	0.3682	0.007055	52.19	<2e-16
AIC 1433973				

- Each coefficient is significant;
- But AIC is too large

Log-Population as an offset variable

There is a strong linear relationship between the scatterplot of log(Car_nbrs) and log(Population). Set log(Population) as offset.

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.974	0.05786	-103.248	<2e-16 ***
log(Apartments)	-0.03473	0.003242	-10.713	<2e-16 ***
log(Builton)	-0.06456	0.001755	-36.782	<2e-16 ***
log(Higheds)	0.02421	0.002596	9.325	<2e-16 ***
log(Seniors)	0.4252	0.007128	59.658	<2e-16 ***
Transit	-0.0001171	5.502e-05	-2.129	0.0332 *
Urban	0.001058	8.677e-05	12.192	<2e-16 ***
log(Vehicles)	0.5714	0.004767	119.869	<2e-16 ***
log(Children)	0.3956	0.01293	30.593	<2e-16 ***
log(Income)	-0.1983	0.01008	-19.681	<2e-16 ***
log(GRP)	0.03015	0.001966	15.338	<2e-16 ***
Fertility	-0.06831	0.005682	-12.021	<2e-16 ***
Persperhh	0.1611	0.007938	20.291	<2e-16 ***
AIC 14223				

- Each coefficient is significant;
- AIC is much smaller now, indicating that set log(Population) as offset is valid.

LUND UNIVERSITY

Fit with log-Population

Now, we set log (population) as the explanatory variable and fit it again to test the hypothesis that there is a proportional relationship between log(Cars_nbr) and log(Population).

Variable	(Intercept)	log(Apartments)	log(Builton)	log(Higheds)	log(Seniors)	Transit	Urban
Estimate	-5.562345	-0.02552618	-0.06789454	0.0409361	0.4157899	3.027331e-05	0.0009199815
Variable	**log(Vehicles)**	**log(Children)**	**log(Income)**	**log(GRP)**	**Fertility**	**Persperhh**	**log(Population)**
Estimate	0.5545457	0.3788941	-0.2290189	0.04022064	-0.05852542	0.1607407	0.9826574

β is close to 1. The hypothesis holds.

Multicollinearity

Now check the VIF value to see if there is multicollinearity.

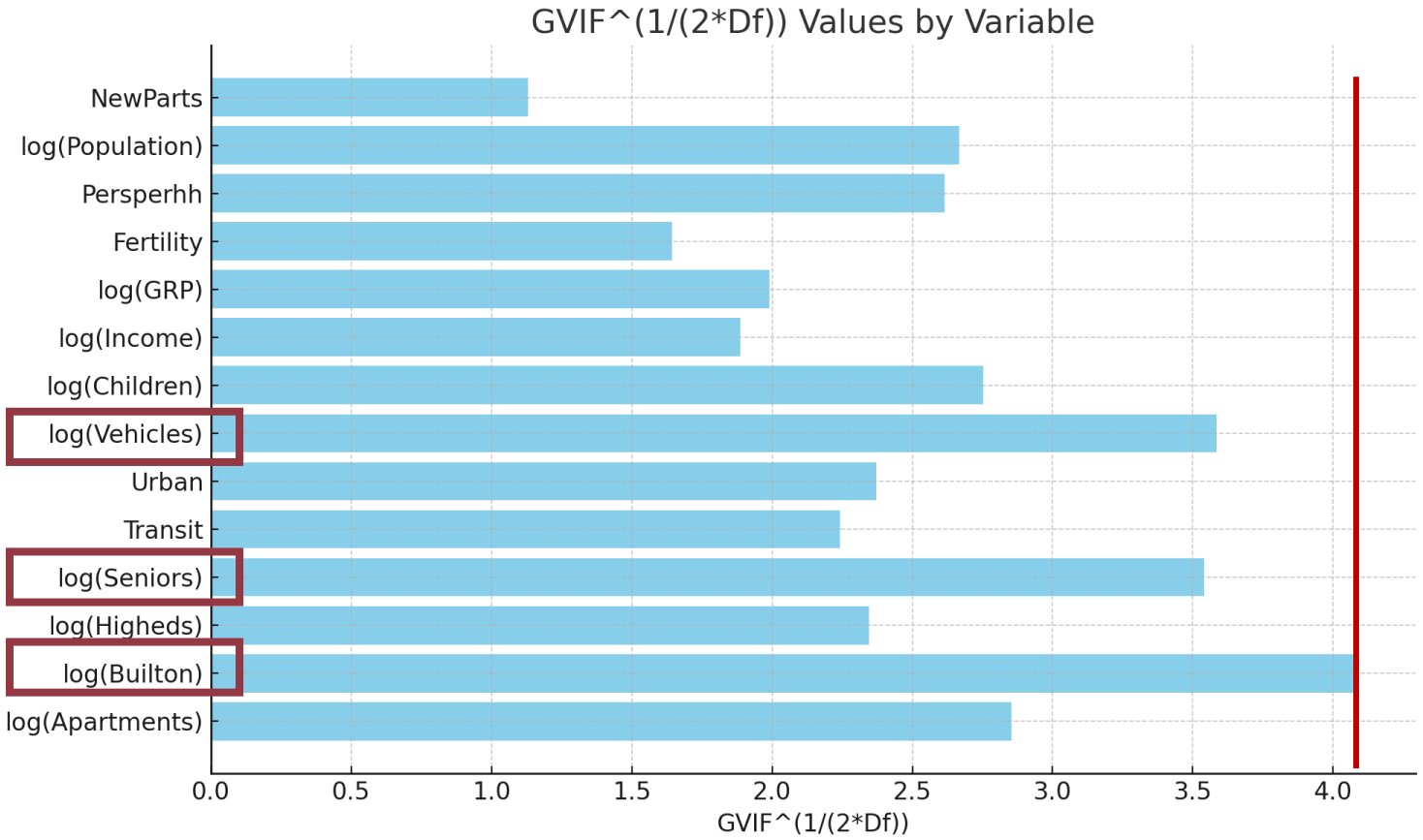
Variable	log(Apartments)	log(Builton)	log(Higheds)	log(Seniors)	Transit	Urban	log(Vehicles)
VIF	7.848394	16.370932	5.489355	11.809884	5.024963	5.641488	12.105274
Variable	log(Children)	log(Income)	log(GRP)	Fertility	Persperhh	log(Population)	
VIF	7.356341	3.522951	3.915096	2.62058	6.671355	6.963236	

Several variables face multicollinearity problems!

LUND UNIVERSITY

Multicollinearity

We introduce “NewParts” in Project1 as a categorical variable to help solve the multicollinearity problem.



LUND UNIVERSITY

AIC and BIC stepwise slection

Up to now our model has been fitted using the full predictors, and we would like to filter out a model that uses fewer parameters and is equally effective with the help of the AIC and BIC.

model	df	AIC	BIC
AIC_stepwise	15	11777.80	11832.85
BIC_stepwise	14	11778.44	11829.82

Compare the BIC_stepwise model with the original one

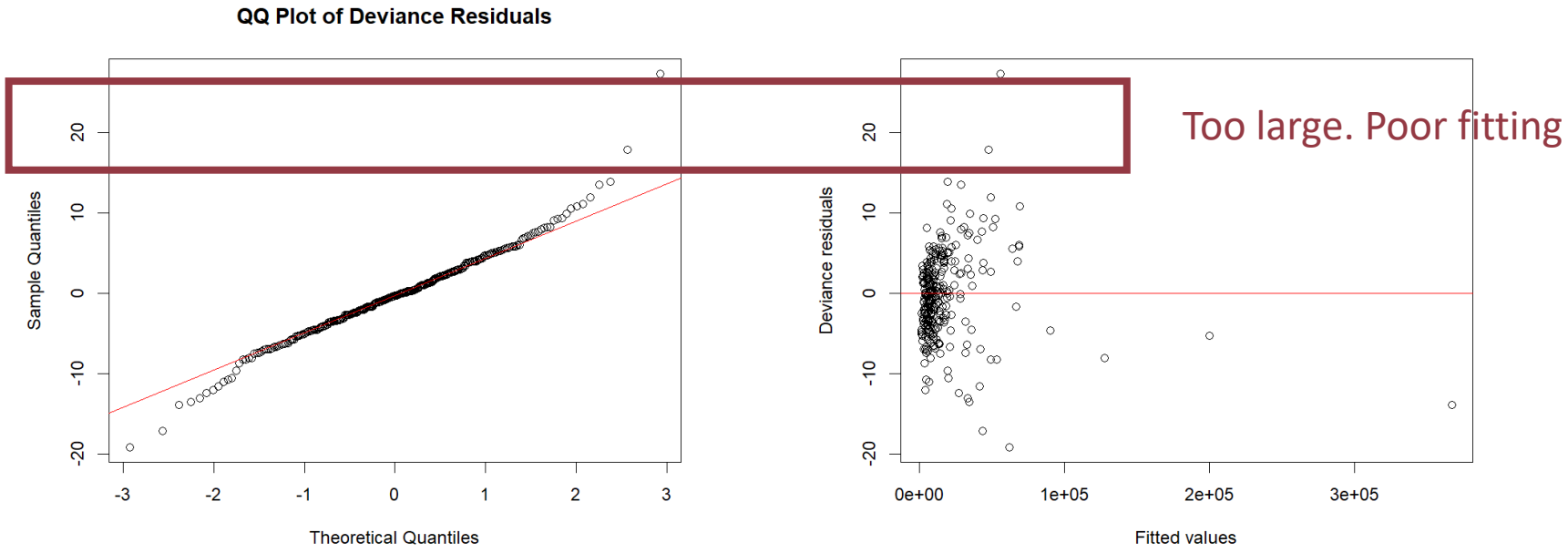
model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Model_full	276.0	8530.3			
BIC_stepwise	274.0	8527.4	2.0	2.8431	0.2413

> 0.05
Acceptable!

Poisson regression optimal model

So we choose BIC_stepwise model as the final model for poisson regression:

$$Cars\_nbr \sim log(Population) + log(Vehicles) + NewParts + Urban + log(Builton) + log(Seniors) + log(Children) + log(GRP) + Persperhh + log(Income) + log(Higheds) + Fertility$$



LUND UNIVERSITY

Poisson regression optimal model

So we choose BIC_stepwise model as the final model for poisson regression:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.842	0.05825	-100.285	< 2e-16 ***
log(Population)	0.9798	0.0008845	1107.662	< 2e-16 ***
log(Vehicles)	0.6068	0.004941	122.818	< 2e-16 ***
NewPartsSvealandandNo	-0.008408	0.001195	-7.036	1.97e-12 ***
NewPartsNorrlandandNo	-0.1103	0.002408	-45.826	< 2e-16 ***
Urban	0.0011	7.891e-05	13.943	< 2e-16 ***
log(Builton)	-0.05644	0.001771	-31.864	< 2e-16 ***
log(Seniors)	0.336	0.007323	45.879	< 2e-16 ***
log(Children)	0.3214	0.01261	25.496	< 2e-16 ***
log(GRP)	0.03434	0.001999	17.178	< 2e-16 ***
Persperhh	0.1371	0.007287	18.822	< 2e-16 ***
log(Income)	-0.1838	0.01018	-18.055	< 2e-16 ***
log(Higheds)	0.04338	0.002685	16.159	< 2e-16 ***
Fertility	-0.03585	0.005384	-6.659	2.76e-11 ***
AIC 11778	Too large. Poor fitting			

Should we use Poisson?

The mean and variance of the data are:

Mean	Variance
17153.48	781627763

Poisson regression assumes that the mean and variance of the target variable are equal, and it is clear that “Cars_nbr” doesn’t fit the Poisson distribution.

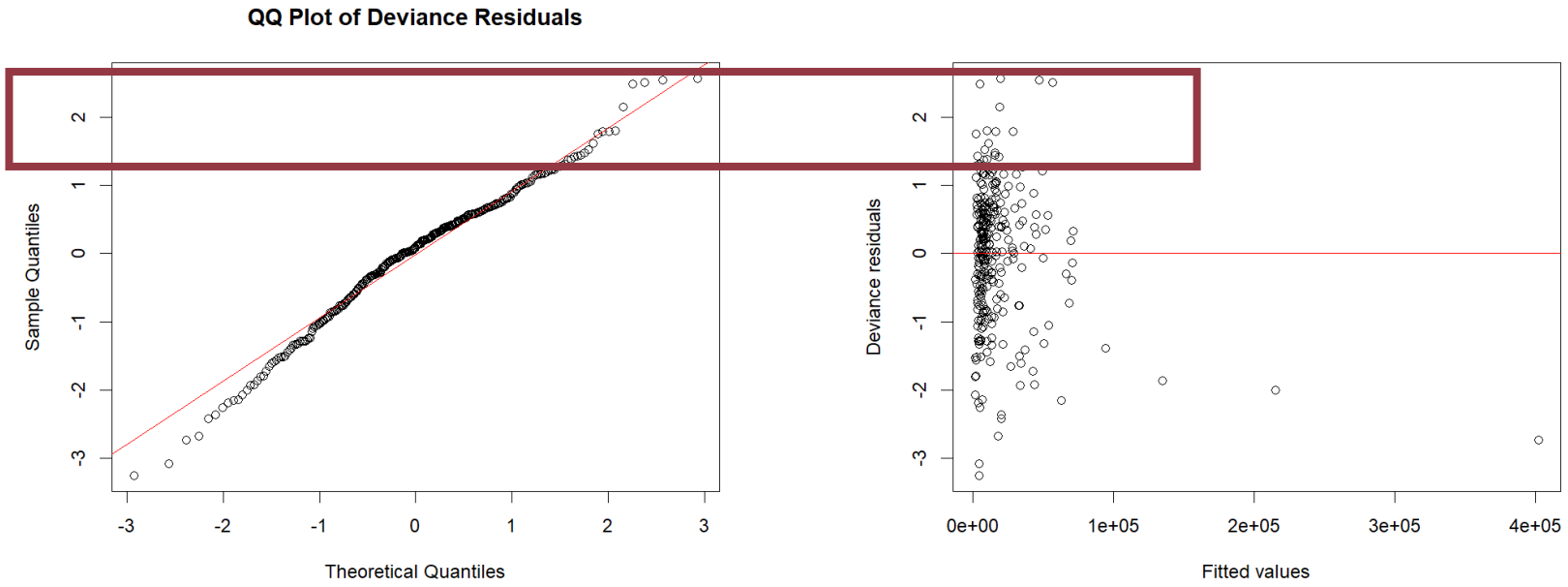
Also, the dispersion test (Pearson Residual Sum of Squares / Residual Degrees of Freedom) result for Poisson regression model is 30.8, which is much larger than 1, indicating a overdispersion problem.

LUND UNIVERSITY

Negative binomial regression optimal model

Similarly, we get the final model for Negative binonial regression:

$$Cars\_nbr \sim log(Population) + log(Vehicles) + NewParts + log(Builton) + log(Seniors) + log(Children)$$



LUND UNIVERSITY

Negative binomial regression optimal model

Similarly, we get the final model for Negative binonial regression:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.504536	0.252474	-21.802	< 2e-16 ***
log(Population)	1.00026	0.0043	232.597	< 2e-16 ***
log(Vehicles)	0.572252	0.025149	22.755	< 2e-16 ***
NewPartsSvealandandNo	-0.005177	0.005947	-0.871	0.384
NewPartsNorrlandandNo	-0.110254	0.010041	-10.98	< 2e-16 ***
log(Builton)	-0.065863	0.009002	-7.316	2.55e-13 ***
log(Seniors)	0.235952	0.030108	7.837	4.62e-15 ***
log(Children)	0.19828	0.037392	5.303	1.14e-07 ***
AIC 4382.4				

Seems better!

Pois vs Negbin - LR Test

We used likelihood ratio test to compare the goodness of fit of two nested models. The negative binomial regression model is an extension of the Poisson regression model because the negative binomial regression model allows for overdispersion.

- Calculate the log-likelihood value for each model
- Calculate the difference in log likelihood values
- Calculate the p-value

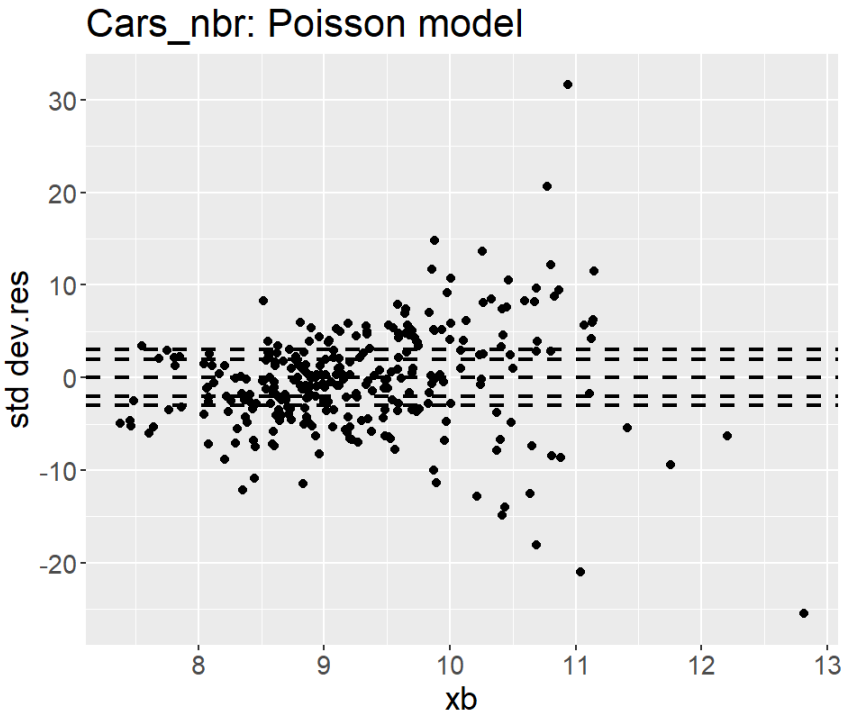
D_pois	D_nb	D_diff	P-value
11750	4364	7386	0

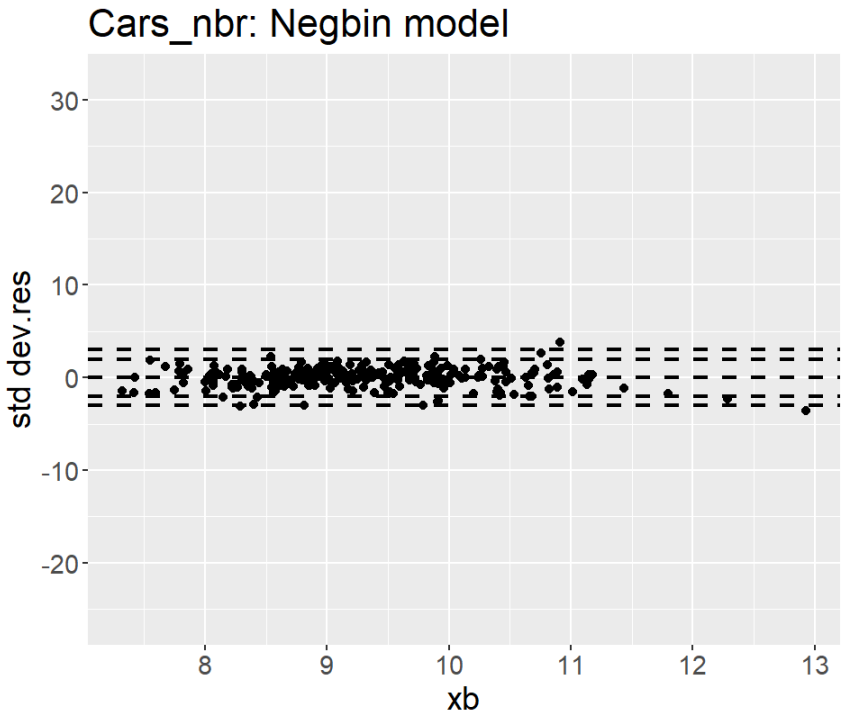
Significant difference in goodness of fit

Complex models significantly outperform simpler models

LUND UNIVERSITY

Pois vs Negbin – std dev.res





	min	max
Poisson	-25.50278	31.63234
Negbin	-3.569294	3.760782

Poisson residuals are too large.
Use negbin model.

LUND UNIVERSITY

Properties of Negbin model

First show the estimated parameters and their confidence intervals:

Variable	beta	2.5 %	97.5 %
(Intercept)	-5.5	-6.0	-5.01
log(Population)	1.0	0.99	1.01
log(Vehicles)	0.57	0.52	0.62
NewPartsSvealandandNo	-0.01	-0.02	0.01
NewPartsNorrlandandNo	-0.11	-0.13	-0.09
log(Builton)	-0.07	-0.08	-0.05
log(Seniors)	0.24	0.18	0.29
log(Children)	0.2	0.12	0.27

- Narrow confidence intervals indicate that the estimates are very precise;
- Confidence intervals for these variables do not contain zeros, indicating that their estimates are reliable at the 95% confidence level.

Properties of Negbin model

We can also check the Adjusted R-squared value:

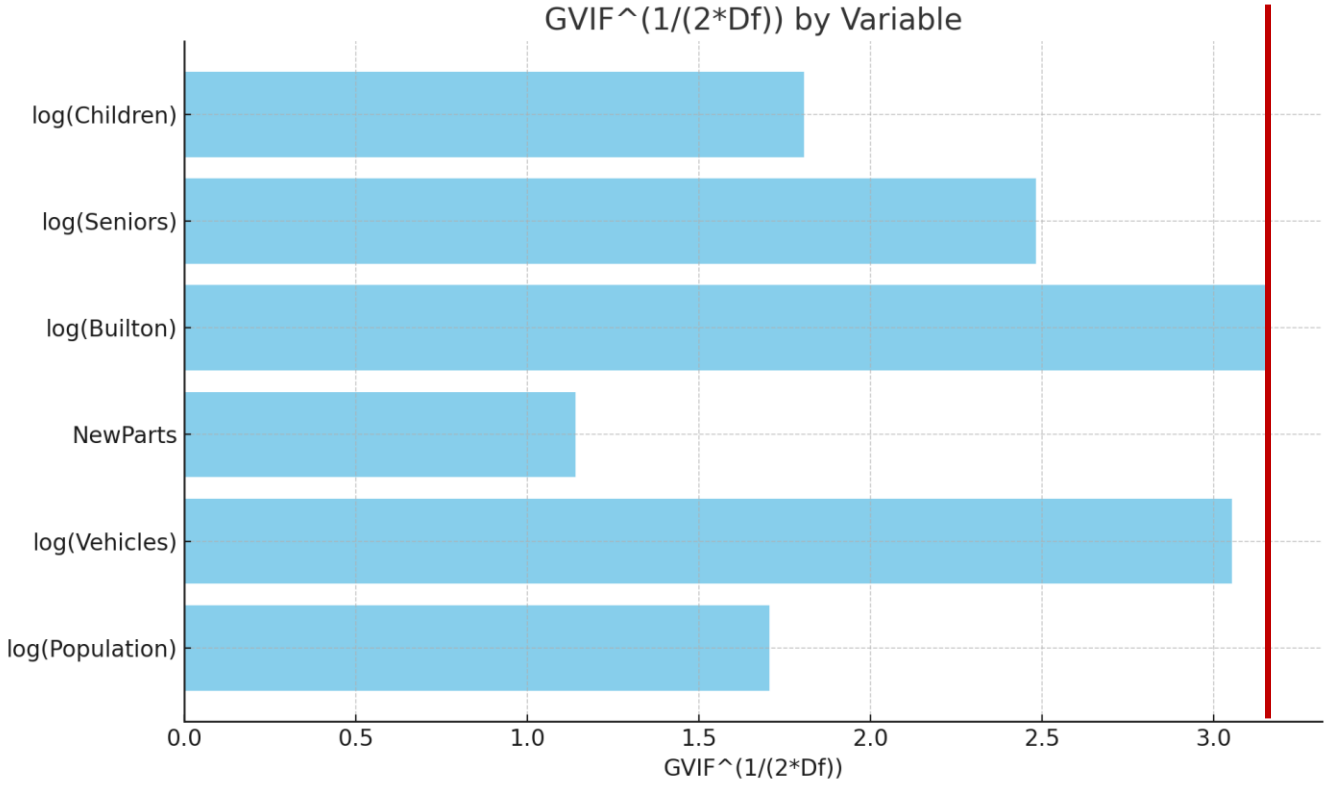
D0	D	p	R2	R2.adj
161386.5	290.9281	7	99.81973	99.81539

- R^2 value close to 1 indicates that the model explains almost all of the variance in the dependent variable. This means that there is a very strong linear relationship between the independent and dependent variables.
- The adjusted R^2 value takes into account the number of independent variables and is still very close to 1, indicating an excellent fit of the model.

LUND UNIVERSITY

Properties of Negbin model

Check the VIF value:

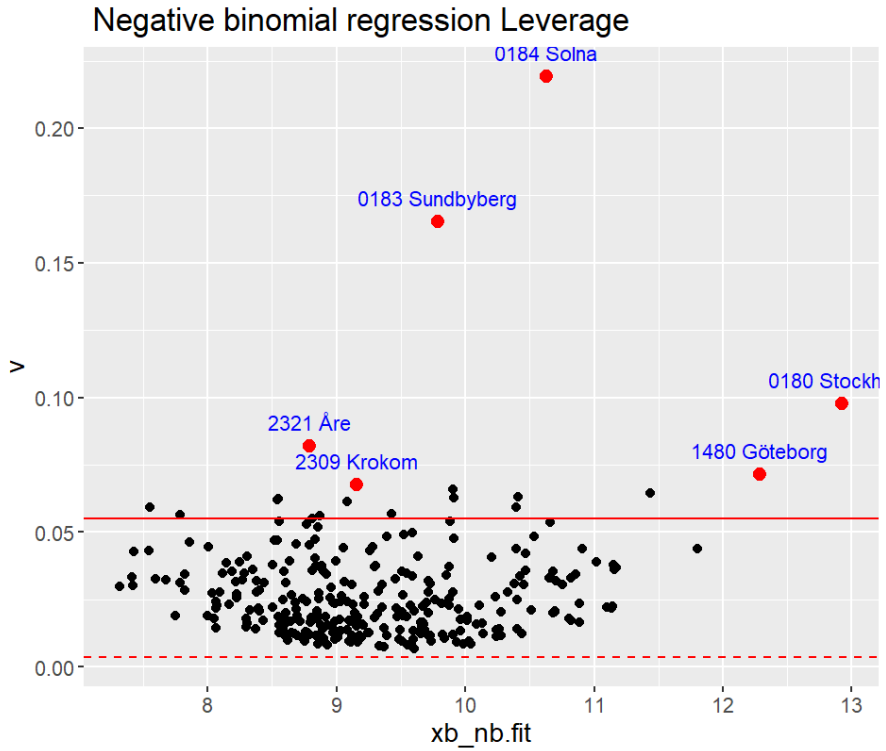


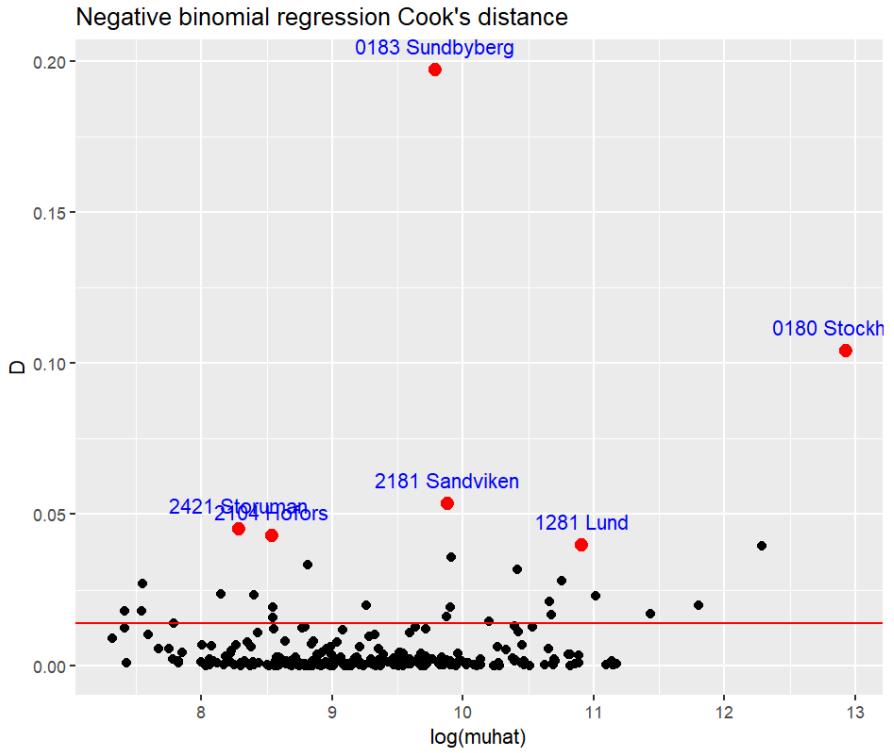
The maximum value is 3.16, which is acceptable.

LUND UNIVERSITY

Properties of Negbin model

Check the data points that have the greatest impact on model fitting through Cook's Distance plot and leverage plot:





LUND UNIVERSITY

LUND
UNIVERSITY