MSG500/MVE190

Linear Statistical Models
Project report

Sharif Zahiri, John Harrysson

14 February 2019

Contents

1 Summary of the minis

First part dedicated to a brief summary of four mini-analysis performed on King's County data set.The data set contains more than 21 thousand observation of 21 variables both numerical and categorical. Objective is to model the price of the houses based on other variables.

The "mini-1" was mostly about data exploration and visual inference from the plots and in "mini-2" a linear multivariate regression model using basically the numerical variables fitted for a sample of 500 randomly selected observations. later on in "mini-3" prediction ability of the previously obtained model were analyzed and the effect of adding the categorical variables to the model were investigated and finally in "mini-4" Poisson and negative-binomial were considered to predict the number of rooms in King's county houses, were we find out COM-Poisson regression may suit better for the data.

1.1 Data review

Plotting variables against each other showed that some of them may correlated which agrees with the calculated Pearson correlation coefficient in figure 3 where there is four pairs with correlation coefficient higher than 0,75. Highly correlated variables may introduce the same information to the model so redundant variable will discard based on previous knowledge and intuition about the model. More over, data overview showed that some variable may need some transformation to facilitate inference for instance Logarithmic transformation used for The Price and some other variables, figure 1. While transformation may ease the inference, multiple implication of complex transformation will make it difficult to interpret the result from the transformed data to the original data. Some variables contain outliers which were excluded from the data set in further analysis and some variable required to be reorganized, such as the variable year_renovated which have been summarized into six classes as shown in figure 2.

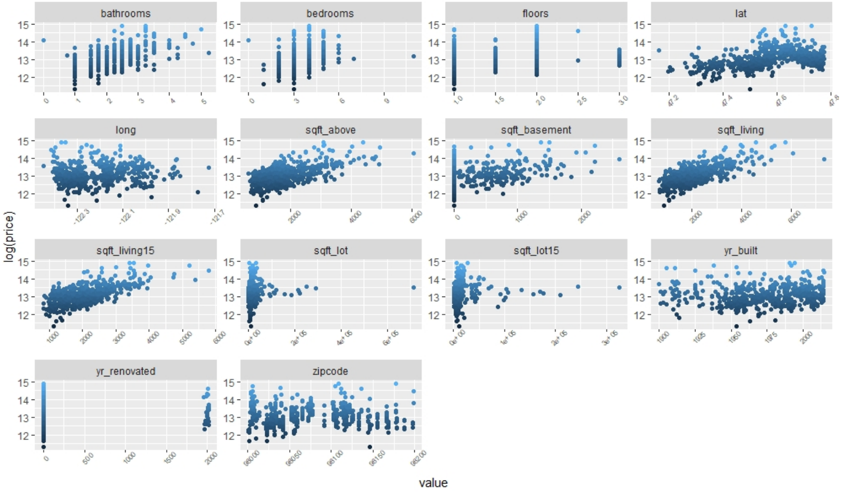


Figure 1: Scatter plot of a sample of 500 observation from few variables against the Log(Price). Some variables show more linearly related to the Log(Price) and some of them, such as Sqft-lot and Sqft-lot15 may need transformation. Note that in Sqft-basement and Yr-renovated it may need some reorganization since large proportion of the points concentrated on zero.

After exploring the data it seems as if $sqft_livinghasalargeimpactandfairlylinearrelationshipontheprice, whichgoeswithir$

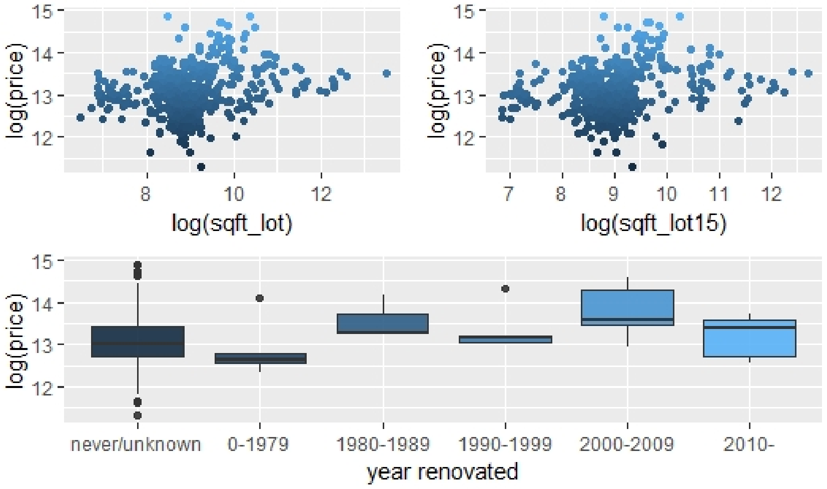


Figure 2: This figure presents the plot of the transformed variables sqft-lot and sqft-log15 and the reorganized version of the variable yr-renovated.

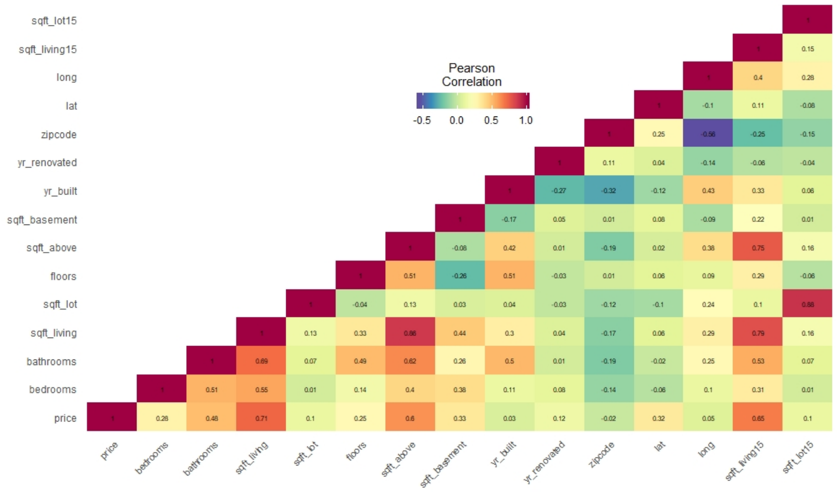


Figure 3: Correlation between variables in data to detect collinearities.

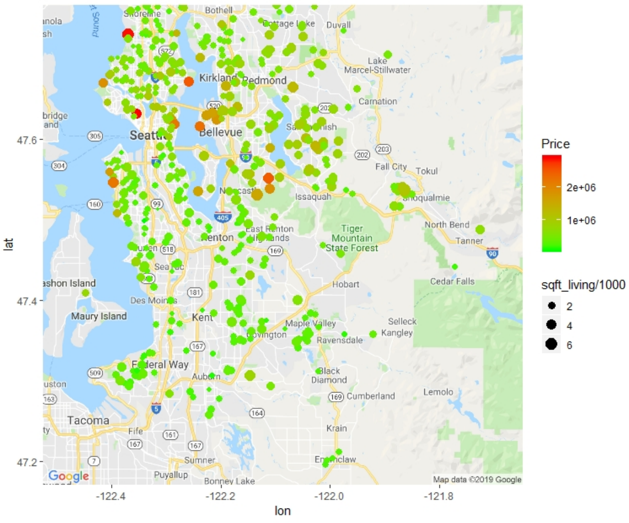


Figure 4: Price plotted on map of King County in relation to sqft_living

1.2 Model fitting

In this section a linear multiple regression model will be constructed for the house Log(price), first with all predictors and afterward we will try to reduce the number of predictors with minimal change in accuracy of the predicted values.

$$y = \beta_0 + \sum_i \beta_i x_i \quad for \quad i \in \{all\ predictors\}$$

Table 8 shows the results from such a model. The term NA in front of the variable $sqft - basement$ *indicates that it needs more investigation, where it possibly denotes some strong collinearity with some other covariate in this case* $sqft{-}living$ *and* $sqft{-}above$ *or more specifically:* $sqft\_living = sqft\_above + sqft\_basement$.

Table 1 presents the summary of the model with out sqft-basement.

Table 1: Regression model summary after removal of *sqft-basement.*Standard deviation in parantheses.

	Dependent variable:
	log(price)
id	0.000 (0.000)
date	0.000** (0.000)
bedrooms	−0.041** (0.017)
bathrooms	0.145*** (0.027)
sqft_living	0.0002*** (0.00004)
sqft_lot	0.050 (0.036)
floors	0.103*** (0.034)
sqft_above	−0.00000 (0.00004)
yr_built	−0.004*** (0.001)
yr_renovated	0.033* (0.018)
zipcode	−0.001* (0.0003)
lat	1.497*** (0.094)
long	−0.338*** (0.126)
sqft_living15	0.0002*** (0.00003)
sqft_lot15	−0.031 (0.039)
Constant	−43.464 (26.956)
Observations	500
R^2	0.755
Adjusted R^2	0.747
Residual Std. Error	0.270 (df = 484)
F Statistic	99.423*** (df = 15; 484)
Note:	*p<0.1; **p<0.05; ***p<0.01

Next to select the variables first backward selection were performed following with forward and step-wise forward variable selection methods, where step-wise forward refer to forward selection with "backward check". The Akaike information criterion (AIC) were utilized to determine which variables to add or remove. The results of the three selection methods can be seen in tables 9. 10 and 11. Interestingly enough, for the current data set all three methods give the exact same model and parameter estimates. Compared to the full model we can observe how the variables that were indicated as significant were also chosen by the selection methods, while the two lot variables were left out.

Apart from step-wise selection there are other model selection methods available, some of the most discussed ones are the LASSO, Ridge and Elastic net, where instead of looking at models and calculating AIC score of the different features after a fit these methods shrink coefficients during estimation. In both LASSO and Ridge the add penalty coefficient will result to more bios but in the other hand will reduce the variation in the final model. The difference is that the LASSO will shrink coefficients to zero while

Ridge will shrink them close to zero but not all the way. The Elastic net regression ,being the blending of two above mention methods, uses a parameter α where $\alpha = 0$ results in Ridge and $\alpha = 1$ results in LASSO regression. Infact it depend on how the Elastec net has been defined, in the package in which it was implemented defines the penalty as $\frac{(1-\alpha)}{2} \parallel \beta \parallel^2 + \alpha \parallel \beta \parallel$.

The Elastic net regression selected to be implemented and parameters α and λ chosen by applying 10-fold cross validation over a grid of values for α and the parameters from the model with the smallest MSE are chosen.

Table 12 shows the result from such a model. Despite the slightly decrease in R^2 model has bin simplified significantly and now it only contains 8 predictors. Figure 5 illustrates some diagnostic plots of this model with the data.

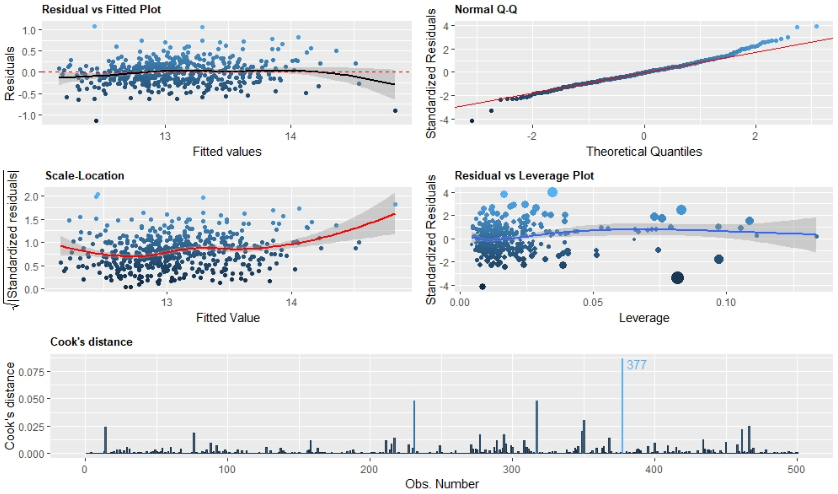


Figure 5: Model diagnostics for Elastic net using lambda with best MSE. The size of dots in the residual vs leverage plot is cook's distance, with large dots having larger cook's distance

1.3 Prediction model for house price

Next the prediction ability of the models were investigated by cross validation, first for the above obtained models and then for the all possible models with different combination of the numerical variables and finally the categorical variables will be included in the analysis.

For the first part, a test data set of 500 observations different from the training set will be selected. Mean squared prediction error (p-MSE) calculated for three different models, first from the step-wise forward method and two from Elastic net method with $\lambda = 0,003$ and $\lambda = 0,02$. The resulting p-MSE's can be found in table 2 and plots of the predicted value vs the actual value can be seen in figure 6.

Table 2: *p-MSE*s for the three different models.

Model:	Elastic.Lambda.1se	Elastic.Lambda.min	Step-wise
p-MSE:	0.083	0.0780	0.0785
Nr. of predictors	8	13	11

From the plots one can observe that in all three models the points are randomly scattered in both side of the 45 degree line and from the p-MSE values one can conclude that the Elastic net model with $\lambda = 0,003$, which is corresponding to the model with the minimum cross validation RSS for training data, has the best performance in predicting with the test data. It is expected since it has the most number of predictors among these three. Looking at the model with the least number of predictors the increase in p-MSE is relatively small.

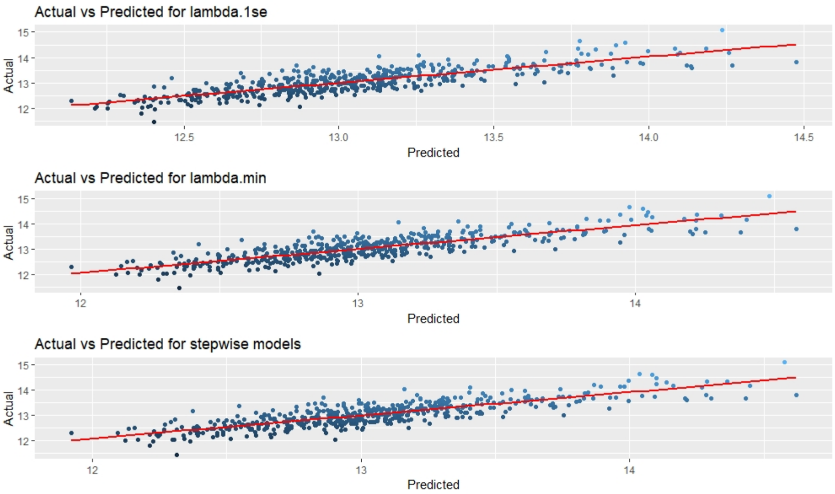


Figure 6: Actual vs Predicted for the three different models

Next, exhaustive model fitting were performed with all possible combinations of the numerical predictors, from the model without any predictor (only the intersection β_0) to the model with all the predictors and as before these models will train with a randomly selected 500 observations and the best model for each model size will selected based on RSS values for the further analysis. For all selected models 10-folded cross validation were performed over entire data set and obtained average p-MSE values for each model presented in figure 7.

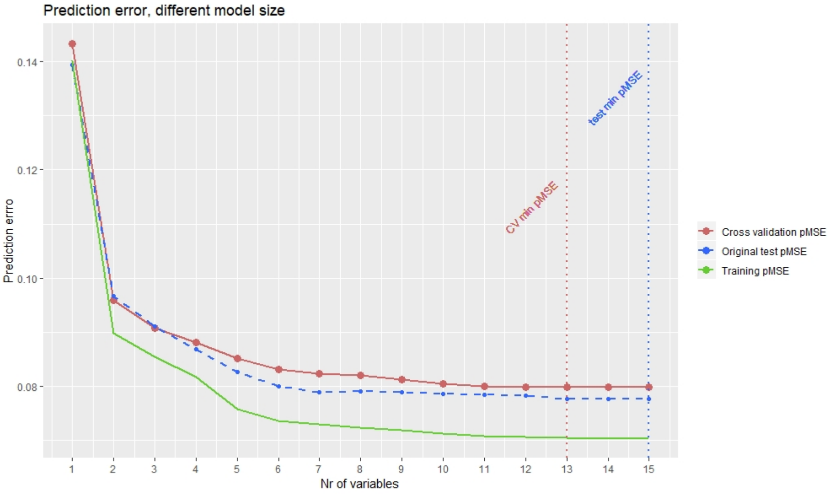


Figure 7: pMSE for best models of each sizes. Both (RSS/n), testing pMSE and CV pMSE are shown with different colors.

Notably despite the fact that in the selected model from Elastic net method only a sample of 500 observation have been used, the p-MSE from that is quite close to the best model with the same size from cross validation over entire data set.

Introducing the categorical predictors to the model makes analysis more complex. Beside the increased model size with adding dummy variables and exponential increase in the number of different possible models, the main concern is when levels of categorical predictors in the data sen contains few observation and it decreases the possibility of them appearing in the training sample data. If the training data does not contain representatives of a certain level of a categorical predictore, it will introduce an extra error to the model and increase the p-MSE simply because the model did not trained for that level and consequently *can not predict that level of the categorical data. However the cross validation method reduces probability of such a scenario where Leave-one-out cross-validation (LOOCV) might seems as a good remedy but the calculation and possessing load specially with a large data set makes it less favorable.*

Before adding the categorical predictors to the model, a brief data review has been performed to detect any obscurity in data and possible interaction between the predictors and some reorganizations has been applied for some of them. The random forest method has been implemented to monitor the possible interaction for six most influential predictors. The result for the longitude parameter raised some concerns about the possible interaction but further analysis did not confirm interaction with any of the other variables. The result fir six variable has been plotted in figure 8.

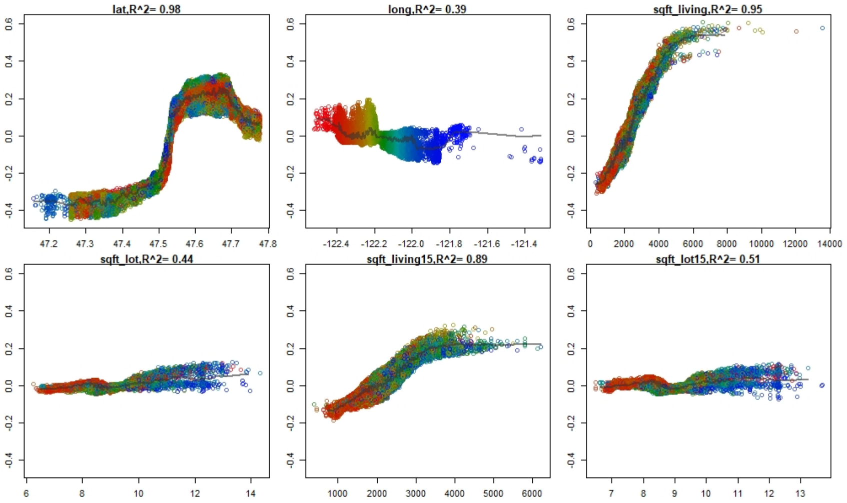


Figure 8: Contribution scores of the 6 most influential variables, colored after gradient of longitude. Distinct color pattern may indicate possible interaction with *longitude*

Adding the categorical predictors to the model requires introducing the dummy variables for every level of each categorical variable. The first level of each categorical variable has been taken as reference level and has been included in the intersection term. The resulted p-MSEs from the 10-folded cross validation over the data set plotted in figure 9. The resulted p-MSEs are improved compared with the model without categorical variables and since there is no penalty for increased complexity in the model it suggests full model with all variables as the best model with minimum p-MSE, but one should consider that after the model with 7 or 8 predictores adding 16 other variables resulted in a small improvement in p-MSE of the model.

1.4 Predicting-model for number of rooms

The forth mini analysis was about constructing a predictive model for the number of bedroom in King County data set. The process was similar to the model construction for the house price and a brief data review conducted and some variables were transformed or reorganized such as the variable bedrooms were limited to 7 and all values larger that 7 replaces with 7 and the variable grade reformatted as a categorical variable with 5 level.

Since number of bedrooms is a count type variable, Poisson regression assumed to be used. Figure 10 shows the histogram of the number of bedrooms in the data set together with Poisson distribution, as it is visible the variation of the data is much less than the Poisson distribution. If the data is over dispersed the negative binomial regression will be more suitable but in this case data were under dispersed and there were not so much text available about a suitable model for this situation.

The ConwayMaxwellPoisson (COMPoisson) being much more flexible distribution were considered to be used instead and as it is visible in figure 10 it fits better to the bedrooms data. The results of the best model obtained by generalized linear model (GLM) is presented in table 3.

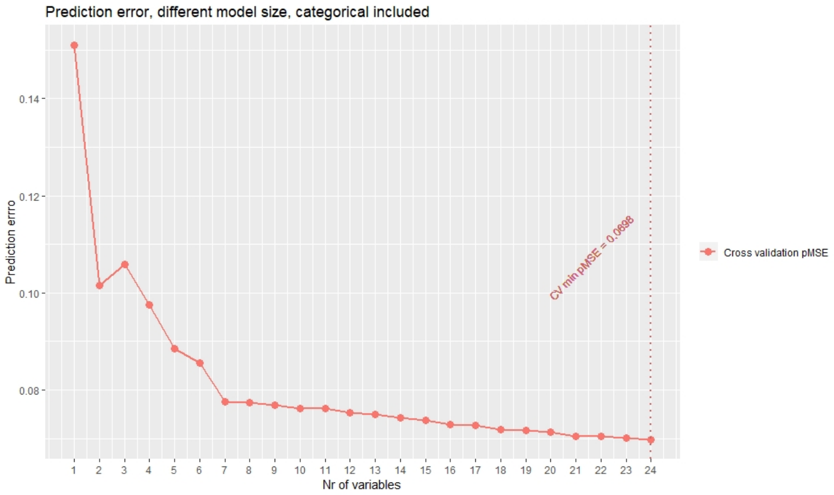


Figure 9: *p-MSE* of cross validation, winning model marked.

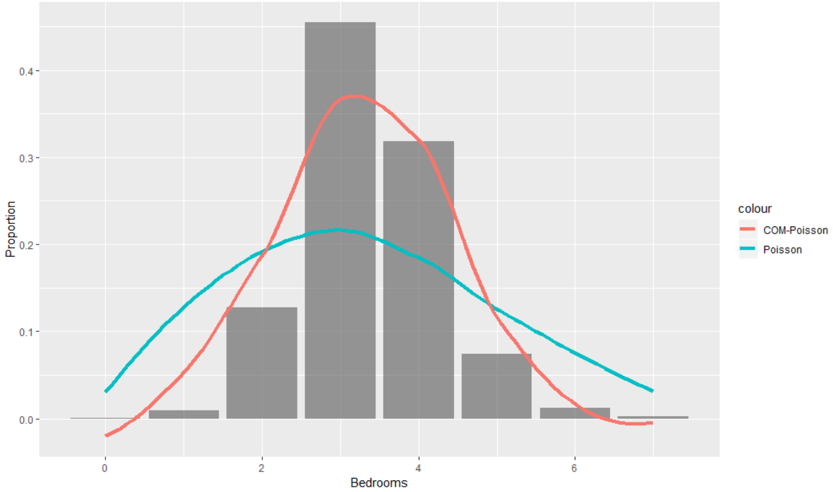


Figure 10: Histogram of the variable *bedrooms* with fitted Poisson distribution. The data variation is less than what expected from the Poisson distribution (under dispersed).

Table 3: Results from the best model obtained by GLM using COM-Poisson.

	Dependentvariable :
	bedrooms
price	-1.2690e-01
bathrooms	2.0627e-01
sqft-living	1.3673e+00
floors	-3.0850e-02
waterfront	-1.6181e-01
view	-5.1082e-02
condition	3.0632e-02
grade	-1.0397e-01
yr-built	-2.2239e-03
zipcode	-6.9269e-05
sqft-lot15	-2.7725e-03
data	Poisson.int
Z	-259.19
P-value	¡2.2e-16
dispersion	0.2400602

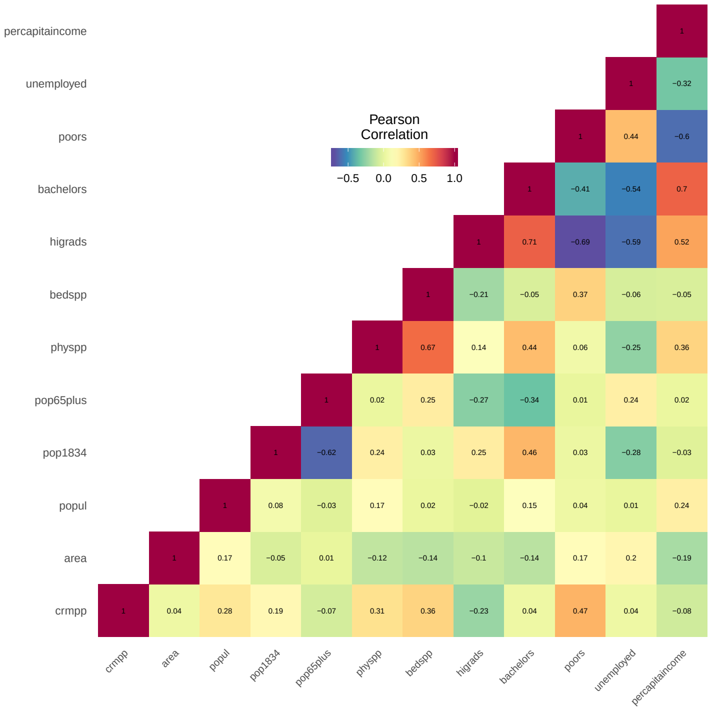
2 CDI analysis

This part of the report dedicated to analyzing the "U.S. county demographic information" data set where the goal is to find the best model to predict the crime per capita rate in U.S. counties.

2.1 Data exploration

U.S. CDI data set consist of 17 variables as columns and 440 entries rows each representing a unique observation. Looking at the variables there is some numerical variables and some categorical variables and a numerating variable "ID". Since we are not interested in the order of the observations and the ID variable does not carry any information about the counties, we can drop it and since the desired response is not in the data set, we calculate the crime per capita times 1000 as our desired response and save it as first column in our data. There is few counties with the same name but from different states present in the data set, we will rename the county names as "county name - state abbreviation" to form a unique name for each different county and set it as rows name in our data set. We will keep the remaining 15 variables for further analysis.

To detect the possible collinearity between numerical variables we calculate the correlation matrix and look for high correlation values both positive and negative. As it is visible in figure 21 there is strong collinearity between few of the variables such as totalincome crime and population. As it is explained in appendix A.1 we decided to drop the variables totalincome and crime from the analysis. Other noticeable collinearity in the correlation matrix was between variables phys and beds, where we assumed that they partially carry same information and finally with changing them to number of physicians per caoita and number of hospital beds per caoita and change their name to physpp and bedspp. Figure 11 shows the final correlation matrix after these adjustments, where there is no longer collinearity between those variables.



percapitaincome												1
unemployed											1	−0.32
poors										1	0.44	−0.6
bachelors									1	−0.41	−0.54	0.7
higrads								1	0.71	−0.69	−0.59	0.52
bedspp							1	−0.21	−0.05	0.37	−0.06	−0.05
physpp						1	0.67	0.14	0.44	0.06	−0.25	0.36
pop65plus					1	0.02	0.25	−0.27	−0.34	0.01	0.24	0.02
pop1834				1	−0.62	0.24	0.03	0.25	0.46	0.03	−0.28	−0.03
popul			1	0.08	−0.03	0.17	0.02	−0.02	0.15	0.04	0.01	0.24
area		1	0.17	−0.05	0.01	−0.12	−0.14	−0.1	−0.14	0.17	0.2	−0.19
crmpp	1	0.04	0.28	0.19	−0.07	0.31	0.36	−0.23	0.04	0.47	0.04	−0.08
	crmpp	area	popul	pop1834	pop65plus	physpp	bedspp	higrads	bachelors	poors	unemployed	percapitaincome

Figure 11: Correlation matrix after removal of crime and totalincome and change of phys and beds to phys and beds per 1000 inhabitants

Next we will investigate the possible linear relation between the dependent variable and independent variables using the scatter plots. Based on resulting plots, figure 22, we decided to transform some of the

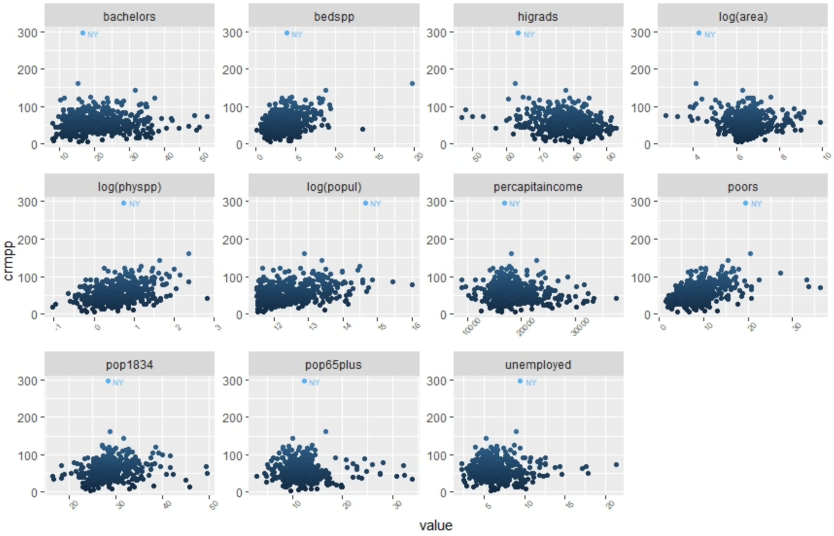


Figure 12: Per capita crime plotted against numerical dependent variables after transforming some of the independent variables. As it is visible in the graphs the scattering of the observation slightly improved compared to the graphs before transformations in figure 22 and the same observation stands out in almost all the plots here as well.

variables to improve their collinearity with the response variable. We applied logarithmic transformation on the variables popul, physpp and area to improve the result with out making them difficult to interpret. Figure 12 shows the plots after transformation. An important observation from the plots was the presence of an extreme outlier. In almost all the plots the same observation from the Kings-NY county had extreme values and most probably presence of this observation in the training data will have strong impact on the final model fitting. We will check this out later.

The variables state and area will be treated as categorical variables. Figure 13 illustrates the box plots of the crime rate in different states at different areas. The same observation from Kings-NY stands out as an extreme outlier here as well. After exploring the data and investigating all the variables we will proceed to model the data.

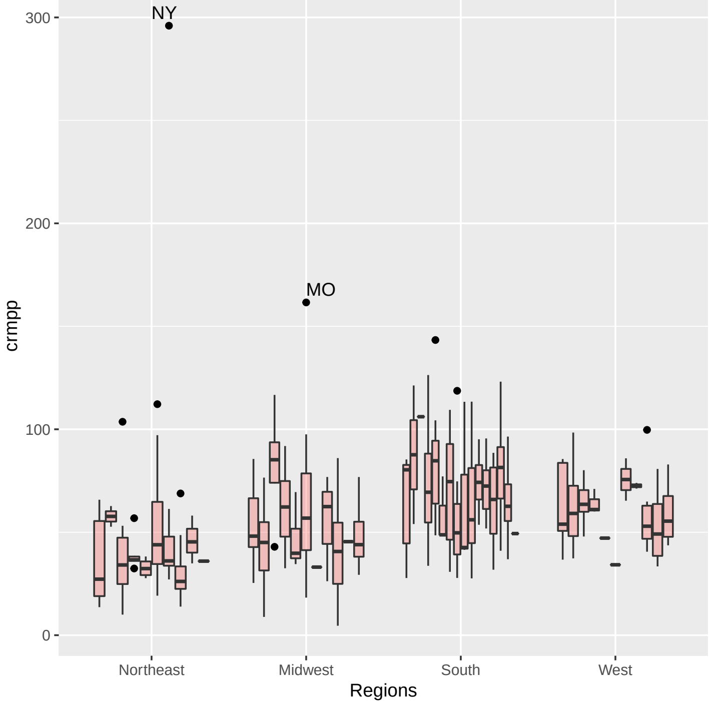


Figure 13: Crime rate in different regions, individual states in respective regions shown as boxplots

2.2 Multiple linear regression model

Figure 14 illustrates average crime rate in each state as a heat map. As it is visible in the map, in general counties from the South area have higher crime rate than other areas and counties from the Northeast area in general have lower crime rate however the most extreme valued outlier was from this area.So based on this holistic general view we will expect our final model to predict lower crime rate value for the counties in the Northeast area and higher rates for the counties from South area. In this section we will try to find a model to predict the crime rate in each county based on CDI data set. For this purpose first we will use the Multiple linear regression method but we need to prepare the data set.

2.2.1 MLR model

The numerical variables are investigated and transformed in previous steps and ready to use for linear regression but categorical variables require introducing the dummy variables. There is two categorical variable in the data set, state and area. Since each area is a set of some states, using dummy variables for each level of these two categorical variables will make each level of area a linear combination of some of the state's levels, so we will drop the state from our data and keep the area. We will set the west area as reference level and use the dummy variables to include the other levels in the model.

First we will tray the model with all the variables. Since we will select the final best model based on their performance on predicting the unseen data, we will randomly select 70% of the data as training data to fit the model and keep the remaining 30% as test data. Table 4 presents the result for the first model With all variables.

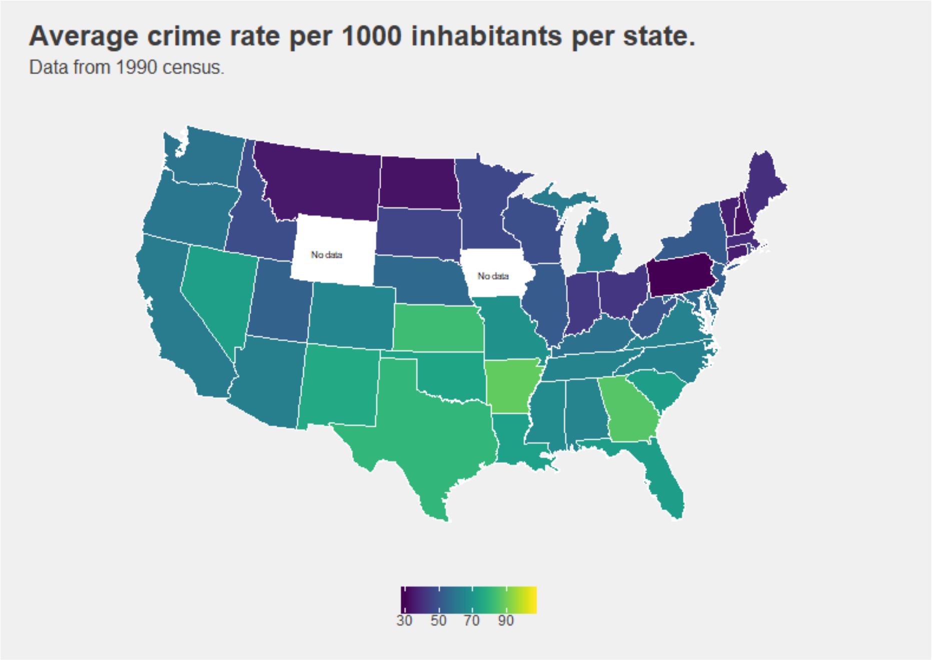


Figure 14: Average crime state-wise. Crime is per 1000 inhabitants, averaged over the counties in the respective state

Table 4: Results for the multiple linear regression model with all variables using 70% of data.

	Dependent variable:
	crmpp
‘log(area)‘	−5.407*** (1.606)
‘log(popul)‘	13.297*** (1.826)
pop1834	0.886* (0.477)
pop65plus	−0.150 (0.430)
‘log(physpp)‘	0.208 (4.313)
bedspp	2.456** (1.008)
higrads	0.371 (0.332)
bachelors	−0.226 (0.407)
poors	2.319*** (0.498)
unemployed	0.755 (0.684)
percapitaincome	0.0003 (0.001)
Northeast	−21.116*** (4.292)
Midwest	−15.405*** (4.245)
South	4.504 (4.023)
Constant	−152.281*** (41.822)
Observations	308
R^2	0.581
Adjusted R^2	0.560
Residual Std. Error	19.224 (df = 293)
F Statistic	28.965*** (df = 14; 293)
Note:	*p<0.1; **p<0.05; ***p<0.01

Variables area, popul, poors, northeast and midwest are significant. Based on their coefficient we can

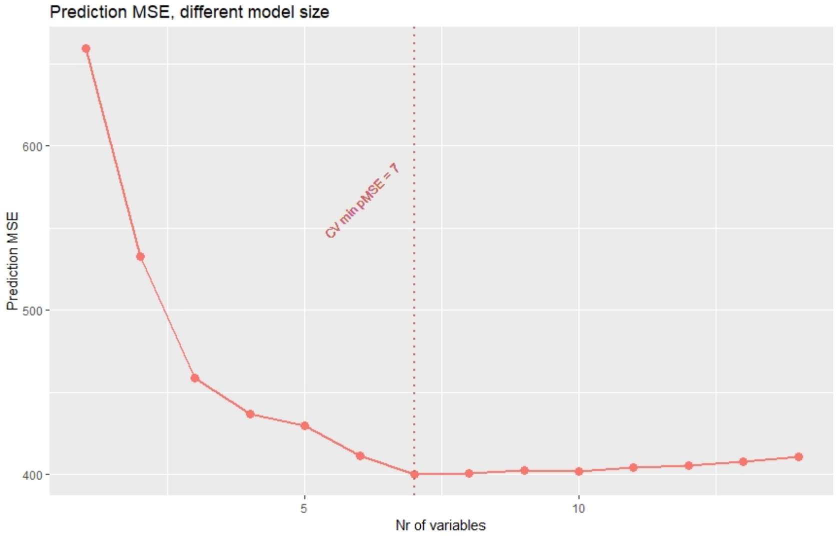


Figure 15: Cross validation error for best models of different sizes.

interpret that increase in the area will decrease the predicted crime rate, increase in popul and poors will increase the predicted crime rate. The variables popul and poors being significant some how makes sense as with more population in a county or more poor individuals in a county one can expect more crime thus higher crime rate but the variable area being significant is not equally straight forward. One hypothesis can be that counties with smaller area has more dense population. For example assuming that if population of a bigger county increased more than a certain level it will for administrative reasons be divided to two counties with smaller areas. Variables northeast and midwest are dummy variables for the levels of the variable area and for interpreting they should be compared to their baseline which is west. It will give us some idea about the model and how different variables will effect the prediction.

Next we will conduct a model selection using exhaustive method. We will train all possible combination of variables for each complexity level from the model without any covariate, all the way to the model with all the covariates. Best performing models for each complexity level will be selected based on prediction mean squared error (p-MSE) using regsubset function in R and performing 10 folded cross validation on training data set. Figure 15 presents the graph of the minimum p-MSE's for each complexity level.

The p-MSE graph has its minimum at complexity level 7, it means that a model with 7 covariates has the over all minimum p-MSE value. Table 5 presents the summary of the best model.

Table 5: Regression summary for winning model with King county NY.

	Dependent variable:
	crmpp
‘log(area)‘	−5.907*** (1.390)
‘log(popul)‘	13.424*** (1.358)
pop1834	0.875*** (0.292)
bedspp	2.298*** (0.612)
poors	2.196*** (0.276)
Northeast	−23.544*** (2.984)
Midwest	−17.250*** (2.868)
Constant	−113.407*** (20.993)
Observations	308
R^2	0.576
Adjusted R^2	0.567
Residual Std. Error	19.090 (df = 300)
F Statistic	58.331*** (df = 7; 300)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.2.2 Investigating the impact of the outliers

In previous step we fined the model that performs the best given the training data, but let recall that we had an extreme value in CDI data set. We selected training data randomly but that extreme observation ended up in our training data and consequently it involved in model fitting process but how big is it's effect on the fitted model? To find out that we will remove that extreme observation from our training data and will run all the model selection steps one more time. The figure 16 presents the p-MSE graph of the best models for each complexity level trained without Kings-NY observation in the training data set.

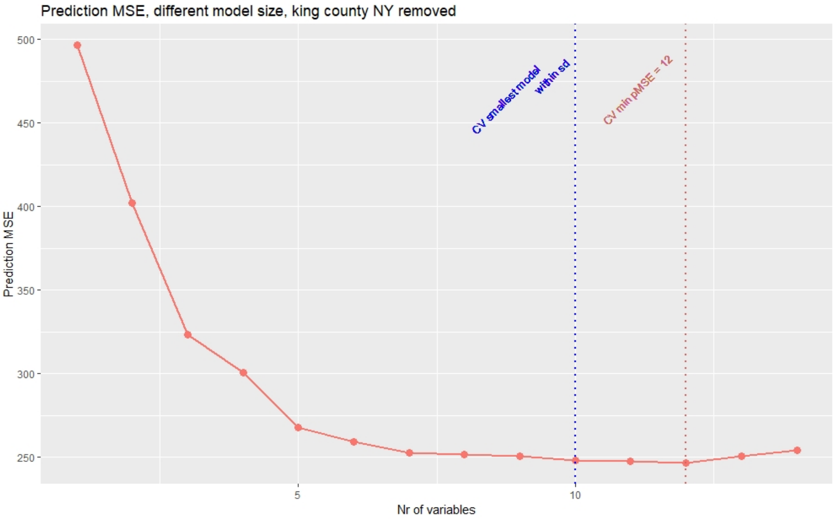


Figure 16: Average crime state-wise. Crime is per 1000 inhabitants, averaged over the counties in the respective state, King county NY removed

In first look we can see that p-MSE values has been reduced significantly such that p-MSE value of the previous best model is larger than almost all models trained without Kings-NY observation. So in general new models perform much better in predicting the observation. The minimum value of the new p-MSE *graph happens at a model with complexity level 12 but the models with less complexity level are also performing reasonably good and they have almost same p-MSE value. We selected the model with 10 covariates as the winning model and table 6 presents the summary of that model.*

Table 6: Regression summary for model without King county NY. Smallest model within one standard error of min pMSE

	Dependent variable:
	crmpp
‘log(area)‘	-2.383^{*} (1.256)
‘log(popul)‘	11.276^{***} (1.119)
pop1834	0.882^{***} (0.257)
bedspp	3.275^{***} (0.514)
higrads	0.332 (0.206)
poors	1.526^{***} (0.329)
unemployed	1.118^{**} (0.530)
Northeast	-22.752^{***} (3.348)
Midwest	-13.551^{***} (3.227)
South	7.478^{**} (3.158)
Constant	-144.652^{***} (24.639)
Observations	307
R^2	0.656
Adjusted R^2	0.645
F Statistic	56.510^{***} (df = 10; 296)
Note:	*p<0.1; **p<0.05; ***p<0.01

Presence of the Kings-NY has a considerably large effect on model selection and the prediction ability of the selected model. But comparing p-MSE value of two model is not a sufficient reason to justify discarding an extreme observation. Figure 17 and figure 16 illustrate some diagnostic plots for respectively the best model with kings-NY observation included in the training data and discard from the training data. Figure 17 shows also that the Kings-NY observation is is an extreme value and beside it rest of the observation behave relay nicely and models has good performance in predicting the response value, but what about unseen data? If the models be used for predicting an other extreme observation they will not preform similarly good. Since the Kings-NY's values does not come from measurement error they may be simply an outcome of the true underlying model and by discarding it we will lose important information about that model. Based on application of the model and purpose of the analysis or nature of the new data set and presence or absence of extreme observations in the data set we can decide to use the earlier model with 7 predictors or later model with 10 predictors.

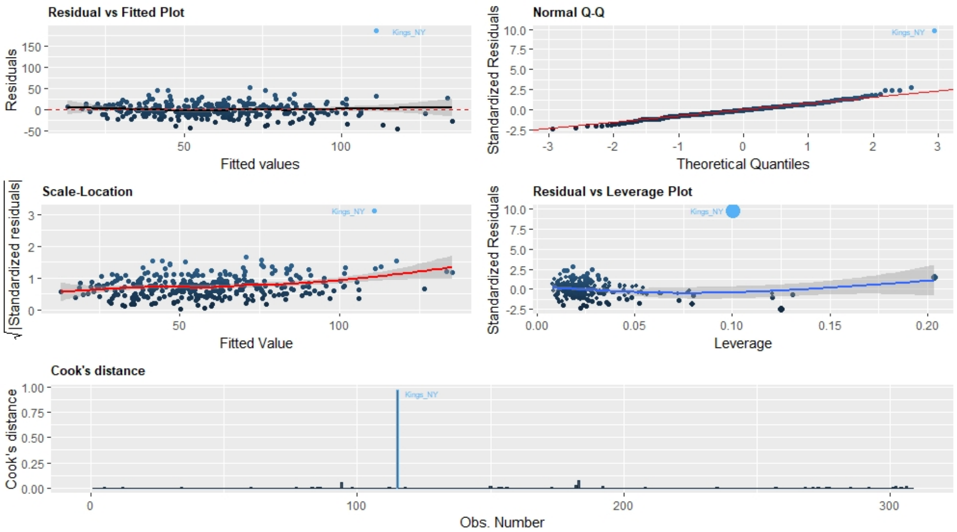


Figure 17: Model diagnostics for final model with Kings county NY included. Size of points in leverage plot represent Cook's distance

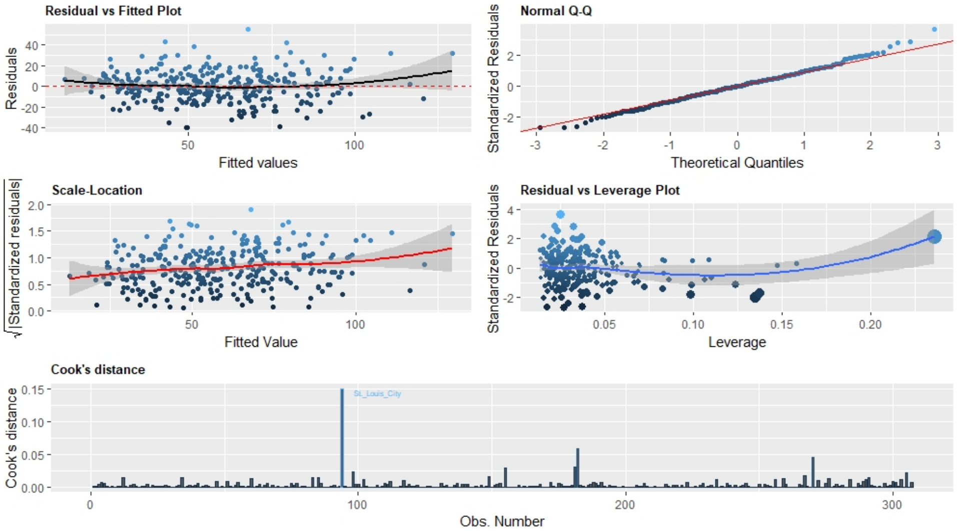


Figure 18: Model diagnostics for final model with Kings county NY removed. Size of points in leverage plot represent Cook's distance

2.3 Poisson/Negative Binomial

Continuing with the poisson modelling we draw upon the experience we got from the normal linear regression. In particular we will reuse the models from the exhaustive search rather than perform an independent model search for the new type of models. The main reason for this is that apart from the normal distribution, other distributions of the exponential family avalible in ***glm*** *are not supported by any dedicated model selection algorithm in R, and writing our own exhaustive search algorithm would be required. While not overly difficult it is not the main goal of this report and we choose to instead use the variables suggested by the normal exhaustive search. Moreover, since we are dealing with count data with a very large rate parameter, approximating it with a normal distribution is not unusual and mathematically supported, so that the models found by the normal exhaustive search shohuld be fairly close to what we would find using a poisson distributed search.*

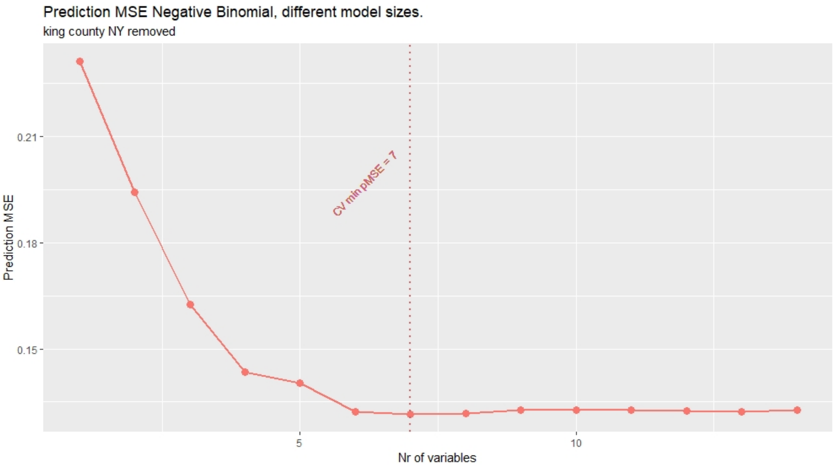


Figure 19: Cross validation error for best models of different sizes. Negative Binomial model

While the poisson distribution is usually used to model count data, we can use it to model the crmpp response by realizing that this is in fact a rate. The count would be represented by the number of crimes and the base, or exposure, would be represented by the population. Given the poisson model assumption we then have

$$\log\left(E\left[Y|X\right]\right) = \beta X \tag{1}$$

And with $Y \propto crime/popul$ we have

$$\log(crime) - \log(popul) = \beta X \Rightarrow \log(crime) = \log(popul) + \beta X \tag{2}$$

*That is, we want to fix the logarithm of our population variable with coefficient 1 to model the rate with the poisson distribution. We do this in R by specifying an offset parameter to **glm**, which in our case is the logarithm of the popul variable.*
Starting by fitting the data to a poisson model consisting of all variables we see immediately that it doesn't fit. In Appendix B, table 13 we can see the resulting summary showing all variables as very significant. This speaks in favour of overdispersion, with the model assuming the conditional variance of the response to be much less than what the data actually suggests. Since the poisson assumption is that the conditional variance is equal to the conditional mean, the observations in the data further away from the mean will pull the fit away from this assumption resulting in the likelihood of these observations under the poisson model being seen as very low, indicating a large significant in the predictors. In Appendix B, figure 23 we can see how the qq-plot against the poisson quantiles shows that most of the data is massed around the tails of the distribution, clearly showing the overdisperion.

To remedy we try instead to fit a Negative Binomial model which mechanically works the same as the poisson model in regards to the offset. We utilize the suggested best models for each size as mentioned earlier and run a 10-fold cross validation with these on the training data, where we use the same training data as in the normal regression with King county NY removed. The resulting pMSE for each model size can be seen in figure 19, where we see that the model with 7 variables has the smallest pMSE. Note that the pMSE is given in log scale of the crime ratio here, in contrast to the normal distributed model from the former section.

In table 7 we see the summary from the winning model. If we compare it to the normal winning model from the same training set in table 6, we notice that the Negative Binomial model fares better in terms *of predictive capacity when we discard higrads, unemployed and South than the normal model did, although by a small margin. Note that this does not mean that the Negative Binomial model has a better predictive capacity overall in comparison to the Normal model, but only within it's own group of models in comparison to the Normal models own group of models.*

If we look at the coefficient estimates themselves and compare to the normal model we see that they are very similar. First acknowledging that the Negative Binomial model is a multiplicative model in comparison to the additive normal model. If we look at the variable log(popul) for example, we see that the expected crime rate increase with 11 units for each unit increase in log(popul) in the normal case and that the expected crime rate increase with 12% for each one unit increase in log(popul) in the Negative Binomial case. This of course means that using the Negative Binomial model we expect the crime rate to increase more if there is already a high crime rate and vice versa. This might seem counter intuitive and the question is how well the model actually fit.

Table 7: Summary for winning model of Negative Binomial regression, coefficient estimates given on log scale.

	Dependent variable:
	crmpp
‘log(area)‘	-0.048^{**} (0.023)
‘log(popul)‘	0.194^{***} (0.023)
pop1834	0.017^{***} (0.005)
bedspp	0.045^{***} (0.010)
poors	0.026^{***} (0.005)
Northeast	-0.526^{***} (0.050)
Midwest	-0.319^{***} (0.048)
Constant	1.219^{***} (0.349)
Observations	307
Log Likelihood	-5,142.371
θ	9.965^{***}(0.791)
Akaike Inf. Crit.	10,300.740
Note:	*p<0.1; **p<0.05; ***p<0.01

To see how well the model fits we look at some diagnostics. We confirm that it does indeed do much better than the Poisson model by looking at the residual and qq-plot in figure 20. From the qq-plot we can read that the overdispersion has mostly been dealt with. There is a slight curvy pattern in the plot but not by any alarming rate. Looking at the residual vs fitted values plot we can see some observations still causing a bit of a problem which should indicate that the data is not perfectly captured by the Negative Binomial model. Note however that looking at the residuals from the perspective of a normal model as done here, with the assumption of evenly distributed residuals, usually don't translate well when considering a model such as the Negative Binomial. This due to the distribution of each fitted value changing. However, when having such a large spread as we have here we can, as mentioned earlier, approximate the distribution by a normal distribution and the residual plot should give us an approximate idea of model fit.

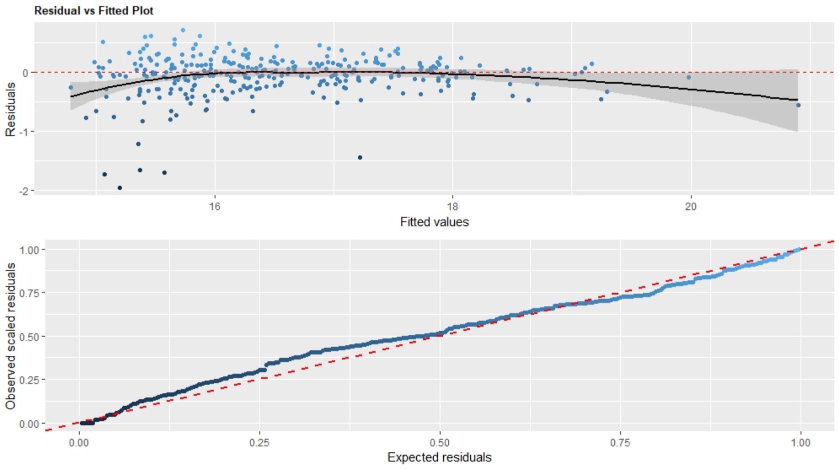


Figure 20: Residual vs fitted and qq-plot for negative binomial winning model. Overdispersion has mostly been handled

2.4 Discussion and conclusion

We have looked at mainly 2 different ways to model the crime rate in the United States, the normal and negative binomial model. Both came up with similar models as end results, with the biggest difference between the two being that one is additive while the other is multiplicative in relation to the conditional mean of the response. This gives two interpretative models of the situation. Which one agrees most with reality is hard to say. While the interpretation of the Negative Binomial may seem counter intuitive at first, with crime generally increasing more based on the predictors if there is already a high crime rate, we have to remember that there may be many underlying factors creating the increase in the predictors. As an example: in the Negative Binomial framework, crime rate increase with two percent for every new poor citizen, meaning an exponentially larger increase in crime rate per poor person in regions with already high crime compared to regions with a current low crime rate. However, there are most likely many underlying factors resulting in this new poor person which would also affect the crime rate. Looking at it from this perspective the model might not seem so outlandish anymore.

Further we also noticed a very significant outlier in King county NY. It differs from other observation in the fact that it has a very high population to area ratio, or in other words a very high density. We choose to show both a model with the outlier and without in the normal distributed case. See former discussion in the text as to why. For the Negative Binomial model we also choose to leave it out. Mainly because we wanted a more general representative model of the crime rate when making comparisons between the normal and Negative binomial models.

Lastly we want to compare the predictive capacity of the two models on the former not seen test data. Predicting the rates we get the following predictive mean square errors:

$$\textit{Normal model:}316.8885\textit{Neg. Binom model:}352.8307 \tag{3}$$

Showing that the normal model is indeed slightly better when it comes to predicting future data.

3 Appendix

A Reasoning behind discarding crime and totalincome

A.1 crime

We have as our dependent variable $Y = 1000 * \frac{crime}{population}$. *It is clear from this that including both crime and popul in the model as independent variables would make* Y *deterministic as given our model assumption:* $Y = E[Y|X] + \epsilon$, *crime, popul* $\in X$. Y *would be measurable with respect to* X *and* $E[Y|X] = Y$, *invalidating the model assumption of normal errors. With this in mind including both crime and popul as independent variables would make any other variables redundant, which is why we choose to remove it.*

A.2 totalincome

The variables of interest are totalincome (totinc), percapitaincome (capinc) and popul. While totinc and capinc are not perfectly collinear they essentially carry the same information, making one redundant. To see this we carry out the following argument:

Let $x_1 = totinc$, $x_2 = capinc = \frac{totinc}{popul}$.

Define

$$< x, y >= \sum_i x_i y_i \tag{4}$$

Then

$$\hat{\beta}_i = \frac{< x_i, y >}{< x_i, x_i >} \tag{5}$$

Now we project x_2 *onto* x_1 *to create an orthogonal basis to calculate the coefficients* $\hat{\beta}$:

$$\gamma_1 = \frac{< x_1, x_2 >}{< x_1, x_1 >} = \sum_i \frac{income_i^2}{income_i^2 popul_i} \approx \frac{1}{popul} \tag{6}$$

Where the subscript indicate that it's the individual incomes and populations for the observations. So we get the residual

$$r_2 = x_2 - \gamma_1 x_1 \approx 0 \tag{7}$$

So that

$$\hat{\beta}_2 = \frac{< r_2, y >}{< r_2, r_2 >} \tag{8}$$

Or in other words, the additional contribution of capinc to the response after totinc has been accounted for, or vice versa, is minimal, leaving us with this very unstable $\hat{\beta}$. *Couple this with the fact that totinc and popul have an almost perfect collinearity and the fact that it doesn't make sense from an interpretative standpoint to say that we fix totinc and popul while varying capinc; it seems reasonable to discard totinc.*

B Extra graphs and tables,

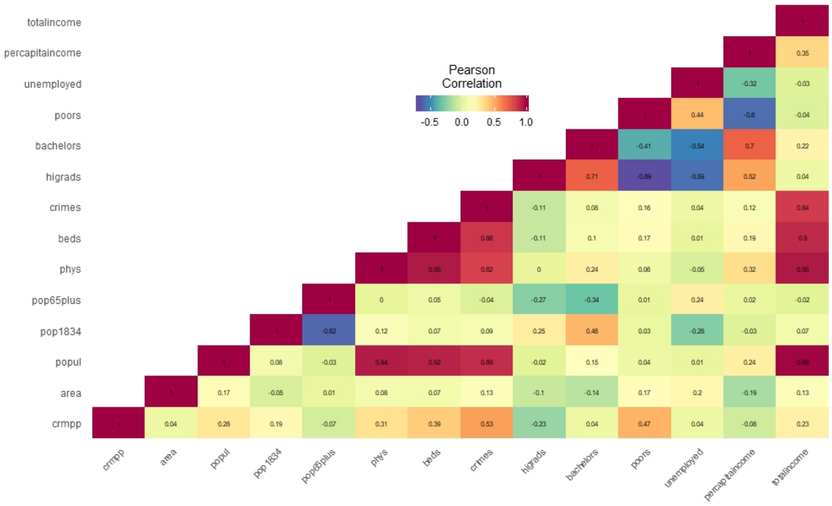


Figure 21: The Correlation matrix of the numerical variables in the data set. Some of the variables appear to be strongly correlated where it can be because of they carrying same information or completely by chance which should be investigated.

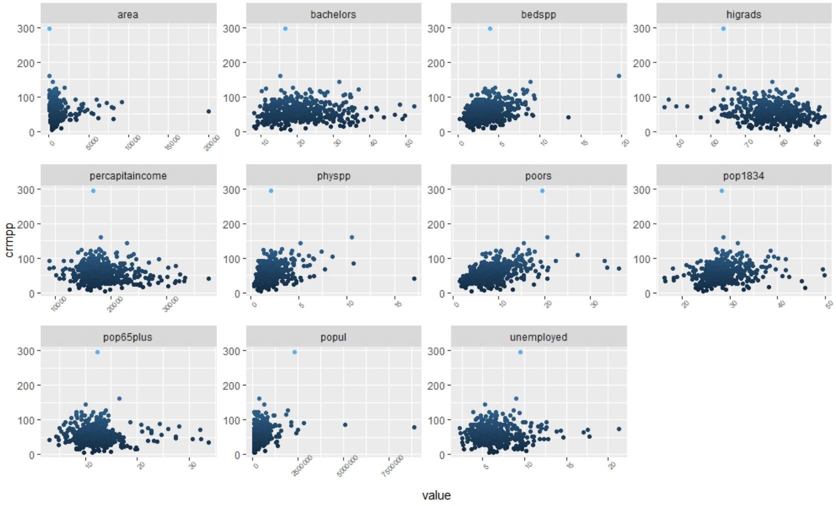


Figure 22: Per capita crime plotted against numerical dependent variables. there is an observation standing out from bulk of the points in the graph possibly an outlier.

Table 8: Summery of regression model with all variables.

	Dependent variable:
	log(price)
id	0.000 (0.000)
date	0.000** (0.000)
bedrooms	−0.041** (0.017)
bathrooms	0.145*** (0.027)
sqft_living	0.0002*** (0.00004)
sqft_lot	0.050 (0.036)
floors	0.103*** (0.034)
sqft_above	−0.00000 (0.00004)
sqft_basement	NA
yr_built	−0.004*** (0.001)
yr_renovated	0.033* (0.018)
zipcode	−0.001* (0.0003)
lat	1.497*** (0.094)
long	−0.338*** (0.126)
sqft_living15	0.0002*** (0.00003)
sqft_lot15	−0.031 (0.039)
Constant	−43.464 (26.956)
Observations	500
R^2	0.755
Adjusted R^2	0.747
Residual Std. Error	0.270 (df = 484)
F Statistic	99.423*** (df = 15; 484)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9: The summary of the selected best model using Backward selection method with AIC as selection criterion.

	Dependent variable:
	log(price)
date	0.000** (0.000)
bedrooms	−0.040** (0.017)
bathrooms	0.143*** (0.026)
sqft_living	0.0002*** (0.00003)
floors	0.091*** (0.028)
yr_built	−0.004*** (0.001)
yr_renovated	0.033* (0.018)
zipcode	−0.001* (0.0003)
lat	1.488*** (0.093)
long	−0.301** (0.120)
sqft_living15	0.0002*** (0.00003)
Constant	−36.382 (26.460)
Observations	500
R^2	0.753
Adjusted R^2	0.748
Residual Std. Error	0.270 (df = 488)
F Statistic	135.545*** (df = 11; 488)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 10: The summary of the selected best model using Backward selection method with AIC as selection criterion.

	Dependent variable:
	log(price)
sqft_living	0.0002*** (0.00003)
lat	1.488*** (0.093)
sqft_living15	0.0002*** (0.00003)
yr_built	−0.004*** (0.001)
bathrooms	0.143*** (0.026)
floors	0.091*** (0.028)
bedrooms	−0.040** (0.017)
date	0.000** (0.000)
long	−0.301** (0.120)
zipcode	−0.001* (0.0003)
yr_renovated	0.033* (0.018)
Constant	−36.382 (26.460)
Observations	500
R^2	0.753
Adjusted R^2	0.748
Residual Std. Error	0.270 (df = 488)
F Statistic	135.545*** (df = 11; 488)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 11: The summary of the selected best model using Step-wise forward selection method with AIC as selection criterion.

	Dependent variable:
	log(price)
sqft_living	0.0002*** (0.00003)
lat	1.488*** (0.093)
sqft_living15	0.0002*** (0.00003)
yr_built	$-$0.004*** (0.001)
bathrooms	0.143*** (0.026)
floors	0.091*** (0.028)
bedrooms	$-$0.040** (0.017)
date	0.000** (0.000)
long	$-$0.301** (0.120)
zipcode	$-$0.001* (0.0003)
yr_renovated	0.033* (0.018)
Constant	$-$36.382 (26.460)
Observations	500
R^2	0.753
Adjusted R^2	0.748
Residual Std. Error	0.270 (df = 488)
F Statistic	135.545*** (df = 11; 488)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 12: Elastic net regularization, chosen model. Using lambda.min.

	$Dependentvariable:$
	bedrooms
id	2.3813e-12
date	203536e-09
bedrooms	-3.1260e-02
bathrooms	1.3772e-01
sqft-living	2.3310e-04
log(sqft-lot)	1.9517e-02
floors	9.4834e-02
yr-built	-3.5353e-03
yr-renovated	3.1852e-02
zipcode	-3.9709e-04
lat	1.4770e+00
long	-2.9313e-01
sqft-lot15	1.7922e-04
Observations	500
Adjusted R^2	0.753
Note:	$\lambda = 0.003; \alpha = 1$

Table 13: Summary of Poisson model utilizing all variables. As seen everything shows up as significant due to the massive overdispersion convoluting the p-values.

	Dependent variable:
	crmpp
'log(area)'	0.032*** (0.0005)
pop1834	0.030*** (0.0002)
pop65plus	0.006*** (0.0002)
'log(physpp)'	0.242*** (0.001)
bedspp	0.005*** (0.0003)
higrads	0.009*** (0.0001)
bachelors	−0.018*** (0.0001)
poors	0.034*** (0.0002)
unemployed	0.015*** (0.0003)
percapitaincome	0.00002*** (0.00000)
Northeast	−0.386*** (0.001)
Midwest	−0.064*** (0.001)
South	0.226*** (0.001)
Constant	−5.271*** (0.013)
Observations	307
Log Likelihood	-231,323.400
Akaike Inf. Crit.	462,674.800
Note:	*p<0.1; **p<0.05; ***p<0.01

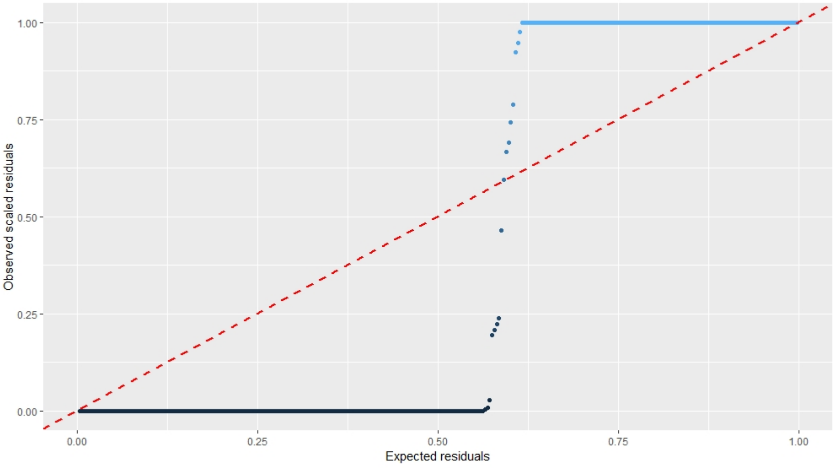


Figure 23: qq-plot of data vs poisson quantiles. The massive overdispersion is seen well as sample quantiles are almost exclusively focused at the tails