Statistical Modeling Based on Model Selection for MLR

Prediction on the Number of Serious Crimes in the United States

by TIANJIAN YANG

March 10, 2019

1 Data Description

1.1 Abstract

Crimes in the United States is a big common concern all the time. Great emphasis is placed and endless efforts have been taken by government to cope with crimes. In this report, we intend to provides some insights into the statistical relationship between the number of serious crimes occurred and basic county demographic information. The dataset of interest in this research is called **CDI**, which provides county demographic information for 440 of the most populous counties in the US, generally pertaining to the years 1990 and 1992. Please refer to **Table 1** for an abstract of this dataset and **Table 2** for the variables within it.

Table 1. Abstract of Dataset – County Demographic Information (CDI).

Data Set Characteristics:	Multivariate	**Number of Instances:**	440	**Source:**	United States Government
Attribute Characteristics:	Numerical, Categorical	**Number of Attributes:**	17	**Copyright:**	Open
Associated Tasks:	Regression, Prediction	**Missing Values?**	No	**Type of Records:**	County

Table 2. Description of Variables in Dataset - CDI.

Variable Name	Description	Format
id	Identification Number	Integer
cty	County name	String
state	Two-letter state abbrev.	String
area	Land area (square miles)	Numeric
pop	Total population (in 1990)	Numeric
pop18	Percent of 1990 population aged 18-34	Numeric
pop65	Percent of 1990 population aged 65 or older	Numeric
docs	Total number of active physicians (in 1990)	Numeric
beds	Total number of hospital beds (in 1990)	Numeric
crimes	Total number of serious crimes in 1990	Numeric
hsgrad	Percent of adult population (25 years and older) who completed 12 or more years of school	Numeric
bagrad	Percent of adult population (25 years and older) with bachelor's degree	Numeric
poverty	Percent of 1990 population with income below poverty level	Numeric
unemp	Percent of 1990 labor force that was unemployed	Numeric

pcincome	Per capita income in 1990 (dollars)	Numeric
totalinc	Total personal income in 1990 (in millions of dollars)	Numeric
region	Geographic region, where 1=northeast, 2=north central, 3=south, 4=west	Categorical

1.2 Objective of Research

Since we have a basic grasp of what information our dataset contains, let’s proceed to clarify our objectives of research on it. Being galvanized by our curiosity along with seriousness, we intend to build multivariate linear regression models to predict the number of serious crimes given the demographic data of a county. R is used as analytic tools across the research.

Ahead of any statistical analysis, the workflow and some key points should be clarified. First, summary statistics are made for training set and test set separately. Second, a thorough pre-processing on data is conducted to ensure the validness of further analysis. Third, in order to obtain our prediction model, we utilize a variety of model selection methods and multiple criterion. Last, RMSE is used for evaluation on the performance of model prediction.

2 Descriptive Statistical Analysis

Let’s subset the raw dataset into variables we are interested in, on which our further analysis will be based. These selected variables and their sample characteristics are summarized in **Table 3**, where a count (percent) table is summarized for categorical variables and ‘mean ± SD’ is calculated for numerical variables. Here, it should be noted that summary statistics is produced for training set (75%) and test set (25%) separately.

Table 3. Sample Characteristics (mean: sample mean, SD: sample standard deviation).
Note: the variables in red color are used to generate response variable *crimesper1000*.

Characteristic	Training Set (n=330)	Test Set (n=110)
area, mean±SD	1031.96±1298.03	1069.76±2142.07
pop, mean±SD	384219.02±597349.57	419386.62±617709.13
pop18, mean±SD	28.61±4.1	28.45±4.47
pop65, mean±SD	12.1±4.01	12.39±3.95
docs, mean±SD	956.8±1755.82	1081.6±1893.12
beds, mean±SD	1415.35±2139.11	1588.46±2696.44
crimes, mean±SD	27164.21±60472.63	26953.85±51203.78
hsgrad, mean±SD	77.28±7.09	78.4±6.75
bagrad, mean±SD	20.86±7.47	21.73±8.19
poverty, mean±SD	8.94±4.88	8.08±3.86
unemp, mean±SD	6.64±2.33	6.45±2.38
pcincome, mean±SD	18405.6±3978.84	19029.13±4275.9
totalinc, mean±SD	7604.08±12729.83	8664.85±13364.92
region	–	–
1=northeast	76 (23.03%)	27 (24.55%)
2=north central	73 (22.12%)	35 (31.82%)
3=south	122 (36.97%)	30 (27.27%)
4=west	59 (17.88%)	18 (16.36%)

The outcome variable, or say response variable, throughout this report and without any transformation on it, is ***crimesper1000*** – the number of serious crimes per 1000 persons. It is derived from ***crimes*** and ***pop*** by the following formula.

$$crimesper1000 = 1000 \times crimes / pop$$

3 Data Pre-Processing

3.1 Regular Pre-Processing

Data Reduction: Raw dataset is pared down into variables we are interested in. Furthermore, ***crimes*** and ***pop*** which are used to generate response variable are also removed.

Dataset Partition: Randomly partitioning dataset into training set (75%) and test set (25%) is done only once in that we assume that the test set is unobtainable during model training.

Missing Values: No missing value is detected in our dataset, so the total 440 observations remain without any deletion.

One-Hot Encoding: For the only categorical variable ***region***, we use One-Hot Encoding method to transform it into dummy matrix – ***region2***, ***region3*** and ***region4***, where the category '1=northeast' is set as reference group. Now we have 15 variables for analysis – one response and 14 predictors.

3.2 Hyper-Parameters

On the one hand, one type of linear transformation – standardization is considered since there are many variables whose data range is significantly large. But the sample size is 440 which is small, so the problem of computational efficiency can be neglected. Besides, the difficulty of model explanation would arise if we conduct standardization.

On the other hand, before doing MLR, whether or not to do non-linear transformation should be checked, in order to improve model fitting. Component-plus-residual plot can help us in this case. In **Figure 1**, we detect a slight non-linear pattern of data points in the subplots of ***pop18***, ***docs***, ***bagrad*** and ***poverty***. Hence, log-transformation is considered. However, after a comprehensive priori experiment, log-transformation improves fitting to training set but degrades the prediction on test set which is our main concern. Arguably, in each subplot, the modeled relationship and LOWESS curve are very close. Also, we would be confronted with the difficulty of seeking for a succinct explanation on our model if we did too many log-transformations. In this reasoning, non-linear transformation on predictors is not needed. Please refer to **Table 4** for the result of our priori experiment.

Table 4. Priori Experiment to Decide the Need or Not for Log-Transformation.
Test Error for Each Model Chosen from Each Model Selection Method.

	A priori	**Best Subset**	**Forward**	**Backward**	**Stepwise**	**Lasso**	**Bivariate p-value**
Test Error without Log-Transformation	17.2289	19.1432	19.1002	19.1432	19.1432	19.0570	20.0230
Test Error with Log-Transformations	17.3256	19.7734	19.6091	20.0120	19.7734	18.8621	21.8454

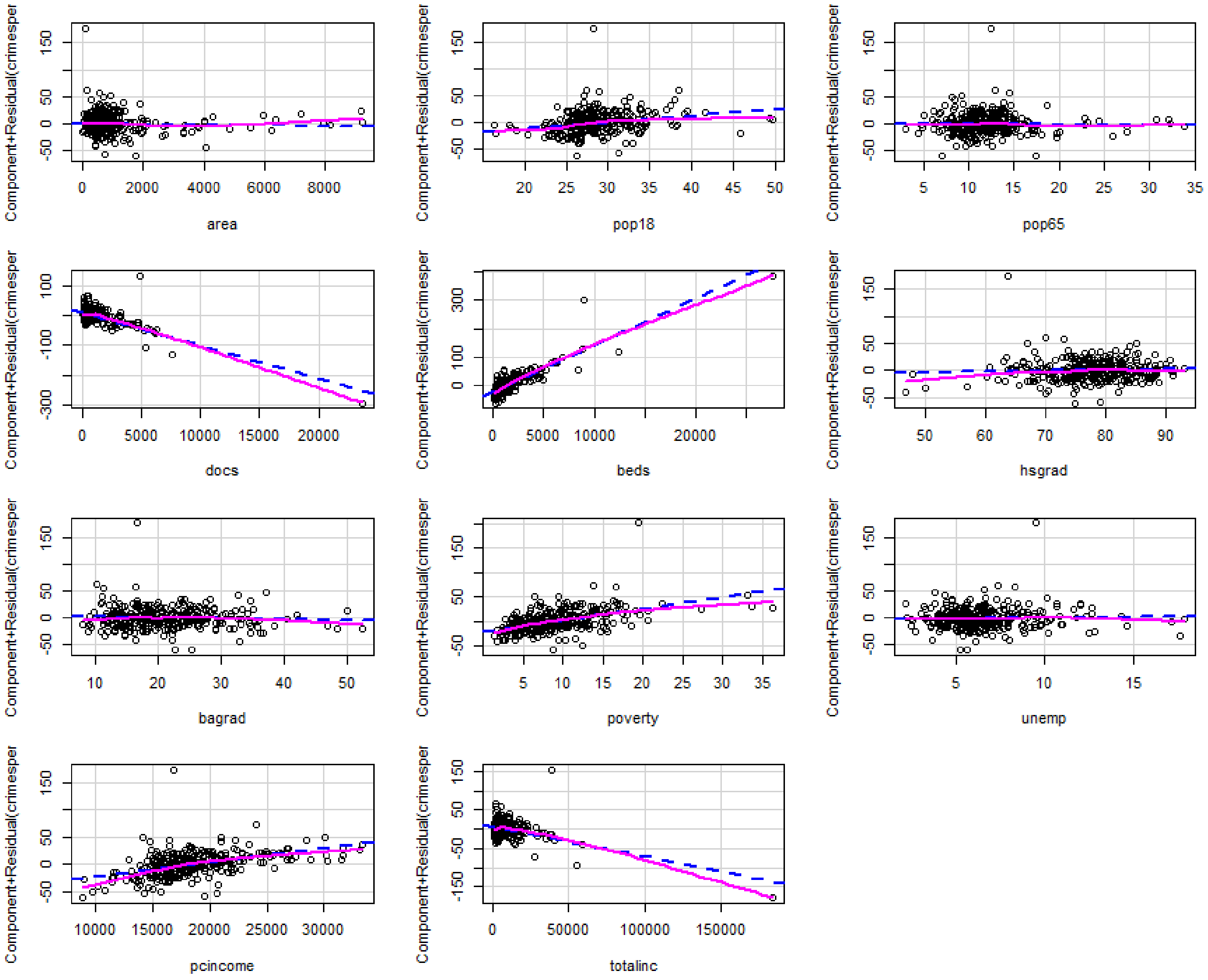


Figure 1. Component-Plus-Residual-Plot Produced from Full MLR Model.
Note: the dash line denotes modeled relationship; the solid line denotes LOWESS curve.

4 Model Selection & Prediction

The MLR models selected from 7 different methods are summarized in **Table 5**. For the sake of clarity, only the estimates of regression coefficients are listed.

In regard to prediction on test set, we need a quantitative evaluation index for the performance of model prediction. Here, we decide to use Root Mean Square Error (RMSE) as the quantitative assessment on each MLR model. In **Table 5**, RMSEs are recorded in the end.

4.1 Best a Priori Judgement

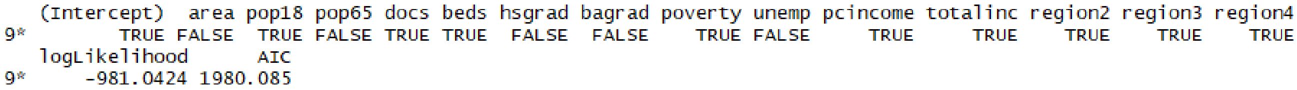
From my personal subjective judgement, empirically speaking, people who know less about laws are more likely to commit crime, so education level is related to crime rates. ***hsgrad*** and ***bagrad*** are selected as predictors. Then, the worse living conditions lead to the higher crime rates. Thus ***poverty***, ***unemp*** and ***pcincome*** are regarded as predictors. Next, the geographic region in US explains crime rates in some degree corresponding to ***region***. Additionally, I reckon the age structure which explains the proportion of the youth, and the size of hospital which implies people's needs, are more or less related to crime rates. It leads us to ***pop18*** and ***beds***. Please see **Table 5** for the estimates of regression coefficients.

Table 5. Regression Coefficients and RMSE for Each Model Chosen from Each Model Selection Method.

	A priori	Best Subset	Forward	Backward	Stepwise	Lasso	Bivariate p-value
***intercept**	-79.6642	-60.3376	-63.8250	-60.3376	-60.3376	-69.8814	-52.6452
area	–	–	–	–	–	-0.0006	–
pop18	1.0408	1.1414	1.2361	1.1414	1.1414	1.1753	0.6277
pop65	–	–	–	–	–	-0.0588	–
docs	–	-0.0113	-0.0111	-0.0113	-0.0113	-0.0109	-0.0063
beds	0.0036	0.0162	0.0161	0.0162	0.0162	0.0161	0.0128
hsgrad	0.4363	–	–	–	–	0.1118	0.7818
bagrad	-0.5805	–	-0.1185	–	–	-0.1398	–
poverty	3.1638	2.3054	2.3073	2.3054	2.3054	2.3397	2.2807
unemp	-0.3434	–	–	–	–	0.2402	–
pcincome	0.0023	0.0024	0.0026	0.0024	0.0024	0.0025	–
totalinc	–	-0.0007	-0.0008	-0.0007	-0.0007	-0.0008	-0.0006
region2	6.8852	8.1226	8.2362	8.1226	8.1226	7.8559	-3.4126
region3	21.9062	22.2983	22.5997	22.2983	22.2983	22.7457	13.0818
region4	13.6093	21.0667	21.4229	21.0667	21.0667	21.8311	–
RMSE on Training Set	20.8649	19.5477	19.5427	19.5477	19.5477	19.5264	21.0090
RMSE on Test Set	17.2289	19.1432	19.1002	19.1432	19.1432	19.0570	20.0230

4.2 Best Subset Selection

```
library(leaps)
library(bestglm)
fit2 <- bestglm(cdi_train[, c(2:15, 1)], IC = "AIC")
fit2$Subsets[which(fit2$Subsets$AIC == min(fit2$Subsets$AIC)),]
```



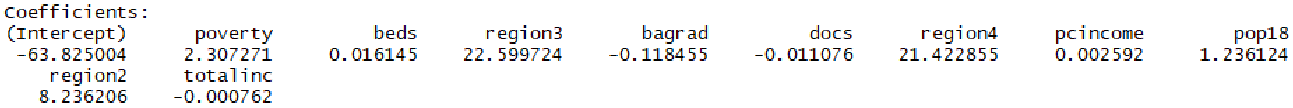
```
   (Intercept)  area pop18 pop65 docs beds hsgrad bagrad poverty unemp pcincome totalinc region2 region3 region4
9*        TRUE FALSE  TRUE FALSE TRUE TRUE  FALSE  FALSE    TRUE FALSE     TRUE     TRUE    TRUE    TRUE    TRUE
   logLikelihood      AIC
9*     -981.0424 1980.085
```

For 14 predictors there are 2^{14} combinations of variables. Based on AIC criteria, MLR models of 9 predictors are concluded as the best. Among them, the model with the least AIC is shown above.

4.3 Forward Selection

```
null <- lm(crimesper1000 ~ 1, cdi_train)
full <- lm(crimesper1000 ~ ., cdi_train)

step(null, scope = list(upper = full, lower = null),
    direction = "forward", trace = TRUE)
```

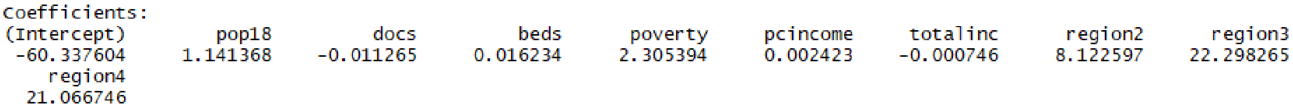


```
Coefficients:
(Intercept)      poverty         beds      region3       bagrad         docs      region4     pcincome        pop18
 -63.825004     2.307271     0.016145    22.599724    -0.118455    -0.011076    21.422855     0.002592     1.236124
    region2     totalinc
   8.236206    -0.000762
```

AIC is used as criteria for our model selection. Please see **Table 5** for more details.

4.4 Backward Selection

```
step(full, scope = list(upper = full, lower = null),
    direction = "backward", trace = TRUE)
```

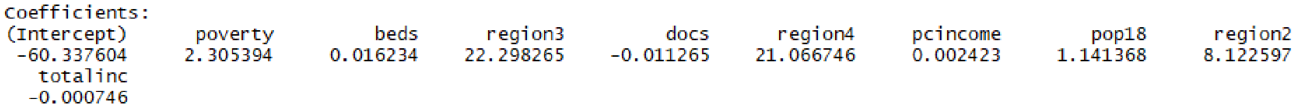


```
Coefficients:
(Intercept)        pop18         docs         beds      poverty     pcincome     totalinc      region2      region3
 -60.337604     1.141368    -0.011265     0.016234     2.305394     0.002423    -0.000746     8.122597    22.298265
    region4
  21.066746
```

AIC is used as criteria for our model selection. Please see **Table 5** for more details.

4.5 Stepwise Selection

```
step(null, scope = list(upper = full, lower = null),
    direction = "both", trace = TRUE)
```



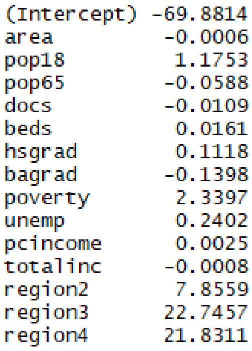
```
Coefficients:
(Intercept)      poverty         beds      region3         docs      region4     pcincome        pop18      region2
 -60.337604     2.305394     0.016234    22.298265    -0.011265    21.066746     0.002423     1.141368     8.122597
   totalinc
  -0.000746
```

AIC is used as criteria for our model selection. Please see **Table 5** for more details.

4.6 Lasso

```
library(glmnet)
set.seed(1234)
cv.lasso <- cv.glmnet(x = data.matrix(cdi_train[, -1]),
                  y = data.matrix(cdi_train$crimesper1000),
                  family = "gaussian", alpha = 1)
plot(cv.lasso)
cv.lasso$lambda.min # 0.01373692

fit6 <- glmnet(x = data.matrix(cdi_train[, -1]),
            y = data.matrix(cdi_train$crimesper1000),
            family = "gaussian", alpha = 1, lambda = 0.01373692)
round(coef(fit6), 4)
```



```
(Intercept) -69.8814
area         -0.0006
pop18         1.1753
pop65        -0.0588
docs         -0.0109
beds          0.0161
hsgrad        0.1118
bagrad       -0.1398
poverty       2.3397
unemp         0.2402
pcincome      0.0025
totalinc     -0.0008
region2       7.8559
region3      22.7457
region4      21.8311
```

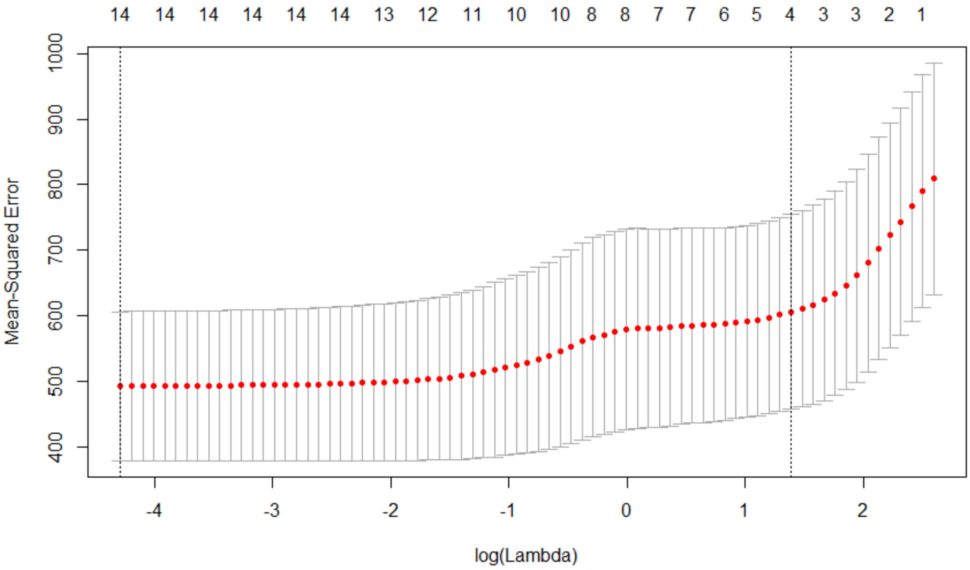


Figure 2. 10-Fold Cross-Validation for Choosing the Optimal Regularization Parameter.

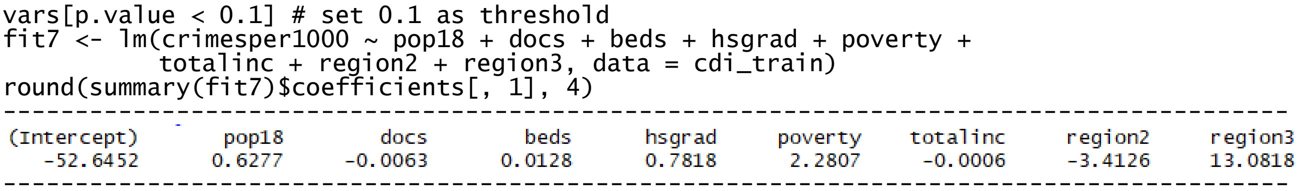
(Mean Squared Errors of MLR Models Under Different Regularization Parameters)

Figure 2 graphically shows the result of 10-fold cross-validation in order to find the optimal regularization parameter. As a result, regularization parameter $\lambda = 0.01373692$ gives us the MLR model with the least MSE.

4.7 Bivariate P-value Threshold Method

```
p.value <- rep(NA, 14)
vars <- names(cdi_train)[-1]
f <- paste("crimesper1000 ~ ", vars, sep = "")

for (k in 1:14) {
  fit0 <- lm(f[k], data = cdi_train)
  p.value[k] <- round(summary(fit0)$coefficient[2,4], 4)
}
```



```
vars[p.value < 0.1] # set 0.1 as threshold
fit7 <- lm(crimesper1000 ~ pop18 + docs + beds + hsgrad + poverty +
           totalinc + region2 + region3, data = cdi_train)
round(summary(fit7)$coefficients[, 1], 4)
------------------------------------------------------------------------------------------
(Intercept)       pop18        docs        beds      hsgrad     poverty    totalinc     region2     region3
   -52.6452      0.6277     -0.0063      0.0128      0.7818      2.2807     -0.0006     -3.4126     13.0818
------------------------------------------------------------------------------------------
```

In this method, we less conservatively set 0.1 as the threshold for p-values.

5 Discussions

5.1 Similarity and Difference

Comparisons between 7 model selection methods except 'Best a Priori Judgement' are summarized into **Table 6**.

Table 6. A General Comparison Between Different Model Selection Methods for MLR.

	Best Subset	**Forward**	**Backward**	**Stepwise**	**Lasso**	**Bivariate p-value**
Computational Efficiency p: # parameters	Low $O(2^p)$	Medium $O((p^2+p)/2)$	High $O(p-1)$	Medium $O((p^2+p)/2)$	Depend on the Number of λ's Tested	High $O(p)$
Model Complexity	Flexible	Flexible	Flexible	Flexible	High (full model)	Depend on Threshold
Hyper-Parameter	Dimension of Subset	No	No	No	L1 Regularization Parameter λ	No
Regularization	No	No	No	No	Yes	No
Resolve Multicollinearity	Not Guaranteed	Yes	Yes	Yes	No	No
Global Optimal Model	Yes	Not Guaranteed	Not Guaranteed	Not Guaranteed	No	No

5.2 Performance of Prediction

Table 7 is truncated from **Table 5**, showing us the performance of prediction for each MLR model selected from 7 different methods. The lower the RMSE on test set is, the more accurate our predictions are, since RMSE is a measure of accuracy to compare the prediction errors of different statistical models on a particular dataset.

Table 7. Training Error and Test Error for Each Model Chosen from Each Model Selection Method.

	A priori	**Best Subset**	**Forward**	**Backward**	**Stepwise**	**Lasso**	**Bivariate p-value**
RMSE on Training Set (Training Error)	20.8649	19.5477	19.5427	19.5477	19.5477	19.5264	21.0090
RMSE on Test Set (Test Error)	17.2289	19.1432	19.1002	19.1432	19.1432	19.0570	20.0230

5.3 A Thinking of Trade-Off

If we were most concerned about using our models to predict future crime rates, the model with the least test error should be implemented – MLR model from 'Best a Priori Judgement'.

However, if we were mainly concerned about having a model with succinct interpretation, some creeds should be followed. First, as simple as possible. Second, variables of lower order that are used to generate high-order terms should be retained. Third, interaction terms regarded as a whole which altogether explain the partial effect of one categorical variable on response should be retained or removed together. Following this reasoning, MLR models from 'Best Subset', 'Backward' and 'Stepwise' can be used for interpretation. Here we note that these three models are exactly the same.

5.4 Limitations

5.4.1 Data Quality

Empirically speaking, first, the amount of data is small since we have only 440 observations. With 75% of it being training set, only 330 instances for training. Meanwhile, test error calculated from the rest 110 instances should not be treated with full confidence. Also, data diversity is limited because the CDI dataset only refers to a specific year 1990. If we intend to use our MLR model, which is learnt from CDI datum of 1990, to predict the crime rates in 2018, it is not guaranteed to be accurate or even totally useless. Moreover, in our dataset, only 12 valid variables can be used for modelling. In fact, more variables carrying extra information may help explain the variability in our response – crime rates. To sum up, a larger and more comprehensive dataset could generally improve the performance of prediction a lot.

5.4.2 Natural Limitations

Theoretically speaking, multivariate linear regression (MLR) model only explains the linear statistical relationship between response and predictors. Probably in nature, the crime rates bear a chaotic and very complicated relationship with these explanatory variables. Perhaps some non-linear or more complex models may have a better performance when making prediction. **Table 8** makes a comparison between our best MLR model and 4 machine learning models, where we see that these more complex models perform slightly better.

Table 8. Comparison of RMSE Between MLR and 4 Machine Learning Models

	MLR	**Neural Network**	**SVM**	**Random Forest**	**XGBoost**
Test Error	17.22893	15.76390	16.34855	16.83793	16.76758

5.4.3 Limitations out of Multiple Objectives

What is weird is that the subjective judgement of a human being results in the best MLR model for making predictions. This phenomenon stems from the fact that we don't make a clear dichotomy between ***Bias*** and ***Variance***. In data science theory, ***Bias*** measures how a model fits to training data whereas ***Variance*** measures a model's capability of generalization on a new dataset. Except for Lasso method which includes regularization, finding the 'best' MLR model is driven by making ***Bias*** as low as possible. However, we do nothing on making ***Variance*** low.

References

Thomas Lumley, Alan Miller (2017) leaps: Regression Subset Selection. https://cran.r-project.org/web/packages/leaps/leaps.pdf. R package version 3.0

A.I. McLeod, Changjiang Xu (2018) Best Subset GLM and Regression Utilities. https://cran.r-project.org/web/packages/bestglm/bestglm.pdf. R package version 0.37

Jerome H. Friedman, Trevor Hastie, Rob Tibshirani (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J STAT SOFTW Vol 33 (2010). DOI: 10.18637/jss.v033.i01

Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li (2005) Applied Linear Statistical Models. McGraw-Hill/Irwin.