200B PROJECT2 REPORT

Xinyang Li

605352032

03/20/2020

➢ Goal and Dataset

This project aims to develop a model to predict the number of serious crimes per capita. Hence, using the CID dataset, the number of crimes is transformed into a crime rate per 1000 persons (crimesper1000). The CDI dataset provides demographic information for 440 counties in the United States pertains to the years 1990 and 1992.

The CDI dataset contains 15 variables, which are IDs of each country (*id*); the land area in square miles (*area*); the total population in 1990 (*pop*); percent of 1990 population aged 18-34 (*pop18*); percent of 1990 population aged 65 or older (*pop65*); total number of hospital active physicians in 1990 (*docs*); total number of hospital beds in 1990 (*beds*); percent of the adult population who completed 12 or more years of school (*hsgrad*); percent of the adult population with bachelor's degree (*bagrad*); percent of 1990 population with income below the poverty level (*poverty*); percent of 1990 labor force that was unemployed (*unemp*); per capita income in dollars in 1990 (*pcincome*); total personal income in millions of dollars in 1990 (*totalinc*); and geographic region (*region*), which is a categorical variable where 1 = northeast, 2 = north central, 3 = south, and 4 = west. To make a better prediction, I added a new variable, population density (*popden*), which is pop/area. Before fitting the model, I first randomly split the dataset into two subsets: training and testing, where there are 330 observations in the training set and 110 observations in the testing set. **Table 1** shows the summary statistics of variables in the two sets separately. In the table, *pop* and *totalinc* have a significant difference in summary statistics between training and testing sets. The mean total population in 1990 is 426251.3 for the training set and is 293289.8 for the testing set. The mean total personal income is 8597.5 millions of dollars for the training set and is 5684.6 millions of dollars for the testing set. There are no other critical differences detected for other variables.

➢ Variable Transformation and Selection

Before conducting model selection methods, I first analyzed if some of the predictors need to do transformations. To see this, I plotted distributions of all predictors individually and fitted loess curved of simple linear regressions of *crimesper1000* (rate of the number of crimes per 1000 persons) on each predictor. After analyzing for all predictors, I found seven predictors that might satisfy the model assumption better after transformation. Those predictors are *area*, *pop*, *docs*, *beds*, *poverty*, *totalinc*, and *popden*. **Figure 1** shows the distributions and loess fits before and after transforming *area*. From the four figures, we can see that the original *area* is not normally distributed, since more than 70 percent of observations have less than 3000 square miles land area but there is an outlier with about 20,000 square miles of land area. This extreme point caused a very poor fit to the data. However, after taking the logarithm of *area*, the distribution of *logarea* is approximately normal and it gave a much better loess curve fit. So the transformation is needed. Similarly, the predictor *pop*, *doc*, *beds*, *poverty*, *totalinc*, and *popden* need to be transformed by taking logarithm for the same reason. **Figure 2 – 7** shows each of the distributions and loess curves before and after transformation.

On the next stage, I conducted seven model selection methods using the training set, which are best a prior judgment, best subset selection, forward selection, backward elimination, stepwise selection, Lasso, and bivariate p-value method by selecting the threshold to be 0.1. The prediction equations for each method were reported in **Table 2**. Finally, after selecting the regression models for the seven selection methods, I used data in the testing set to calculate the test errors using the formula $\sqrt{\frac{1}{110}\sum_{i=1}^{110}(y_i-\hat{y}_i)^2}$, respectively and found the test errors about 30 to 40 for all selection methods. After observing the testing dataset, I found that the largest y_i (*crimesper1000*) is 295.99, which is about 156% larger than the second largest y_i. This outlier dramatically increased the test error so I decided to drop it. After dropping the outlier and calculating the root mean test errors, from the **Table 2** again, we can see that the bivariate p-value method has 14.79 as the smallest test error, and the backward elimination has test error 17.79, which is the largest among all methods. Therefore the bivariate p-value method can predict new observations better than other methods, and the backward elimination may produce the worst prediction on new observations among all methods.

- **Findings and Further Concerns**

From the prediction equations table **(Table 2)**, we can find the forward selection and stepwise selection produced the same prediction equations, with identical parameter estimates. In addition, I also noticed that the bivariate p-value method produced the equation with the most predictors selected: all predictors are selected except *logarea*, *bagrad*, and *unemp*. To conduct the bivariate p-value selection, I regressed every single predictor on *crimesper1000* and selected all predictors with p-value less than 0.1. After running all simple linear regression models, I found that more than half of them have p-value less than 0.0001. Therefore the bivariate p-value method contains the most predictors and has the least test error. Further, I noticed that some of the predictors appeared in all selection methods but some of the others appeared only in a few methods. For example, the categorical variable *region* appeared in all seven models, and the parameter estimates for each dummy variable in the seven methods are quite similar; and *logpoverty* also appeared in all models with similar parameter estimates. On the other hand, the variable *logarea* appeared only in my "best a prior judgment", indicating that my prior judgment might not fit the real condition.

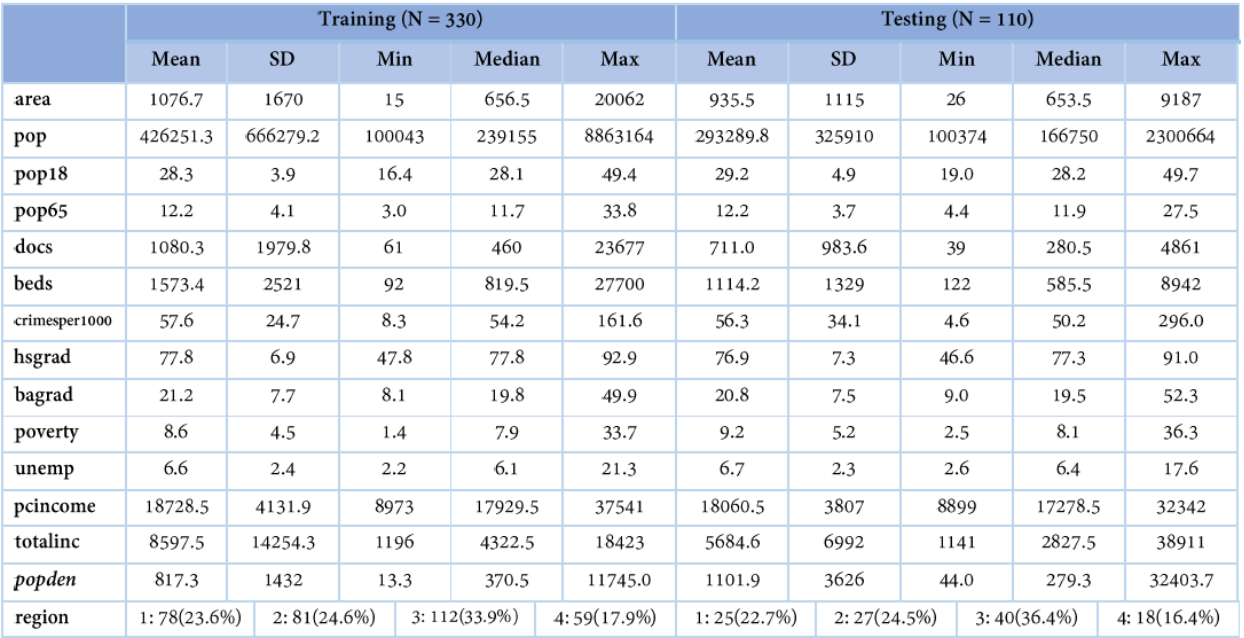
I would choose the best subset selection model as the most interpretable model. This model contains 8 predictors (including one categorical variable) and does not have predictors that appeared only in a few methods such as *logarea*, *unemp*, and *logpopden*. Also, it has a decent test error 15.07, which is the second least test error among all selection methods. Since it does not contain too many predictors, it is easier to interpret the model. For example, for the coefficient of *logpoverty*, it means when the percent of 1990 population with income below the poverty level is doubled, the crime rates per 1000 persons will increase by about 16.8. In addition, From **Table 3**, the parameter estimates table, we can see if a predictor is significant by analyzing its p-value. From the p-values, we know that all predictors are less than 0.05 except *pop65* and *bagrad*, indicating that these two variables do not contribute much to explain the variation of the crime rates per 1000 persons (*crimesper1000*). Also, both *logpop*, *logpoverty*, *logtotalinc*, and *region* have p-values < 0.0001, so they are critical for explaining the variation of crime rates.

Further, I think the bivariate p-value method can give a better prediction for future observations than any other method since it has the smallest test error. With larger test MSE, we may be overfitting the data and we want to avoid this situation. This model involves 10 predictors, which is more than the number of predictors in all other methods. From **Table 4**, we may see that *pop65*, *logdocs*, *logbeds*, *hsgrad*, and *logpopden* do not have significant contributions to explain the variation of crime rates per 1000 persons.

In the end, the final models for predicting crime rates per capita may not be accurate because of some limitations. First, the CDI dataset only provides information that generally pertains to the years 1990 and 1992. Therefore the data collected from 30 years ago may not predict the crime rates today accurately. In addition, in this project, we did not consider the interactions between two variables. If interactions exist, it may not predict the number of serious crimes per capita accurately if we did not detect and address them.

Tables and Figures

Table 1: Summary Statistics



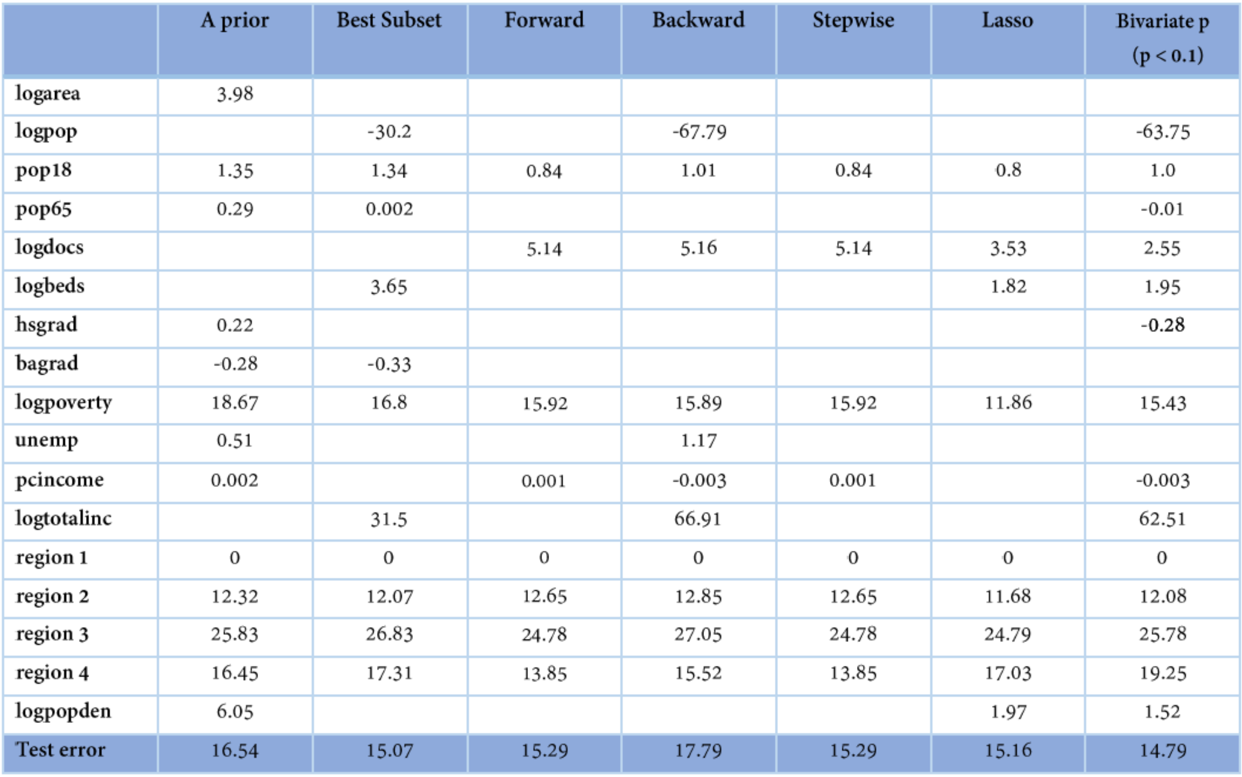
	Training (N = 330)					Testing (N = 110)				
	Mean	SD	Min	Median	Max	Mean	SD	Min	Median	Max
area	1076.7	1670	15	656.5	20062	935.5	1115	26	653.5	9187
pop	426251.3	666279.2	100043	239155	8863164	293289.8	325910	100374	166750	2300664
pop18	28.3	3.9	16.4	28.1	49.4	29.2	4.9	19.0	28.2	49.7
pop65	12.2	4.1	3.0	11.7	33.8	12.2	3.7	4.4	11.9	27.5
docs	1080.3	1979.8	61	460	23677	711.0	983.6	39	280.5	4861
beds	1573.4	2521	92	819.5	27700	1114.2	1329	122	585.5	8942
crimesper1000	57.6	24.7	8.3	54.2	161.6	56.3	34.1	4.6	50.2	296.0
hsgrad	77.8	6.9	47.8	77.8	92.9	76.9	7.3	46.6	77.3	91.0
bagrad	21.2	7.7	8.1	19.8	49.9	20.8	7.5	9.0	19.5	52.3
poverty	8.6	4.5	1.4	7.9	33.7	9.2	5.2	2.5	8.1	36.3
unemp	6.6	2.4	2.2	6.1	21.3	6.7	2.3	2.6	6.4	17.6
pcincome	18728.5	4131.9	8973	17929.5	37541	18060.5	3807	8899	17278.5	32342
totalinc	8597.5	14254.3	1196	4322.5	18423	5684.6	6992	1141	2827.5	38911
popden	817.3	1432	13.3	370.5	11745.0	1101.9	3626	44.0	279.3	32403.7
region	1: 78(23.6%)	2: 81(24.6%)	3: 112(33.9%)	4: 59(17.9%)		1: 25(22.7%)	2: 27(24.5%)	3: 40(36.4%)	4: 18(16.4%)	

*In region, 1 =northeast, 2 = north central, 3 = south, 4 = west.

*crimesper1000 = 1000*crimes/pop, is the crimes rate per 1000 persons.

*popden = pop/area, is the population density.

Table 2: Prediction Equations for Model Selections



	A prior	Best Subset	Forward	Backward	Stepwise	Lasso	Bivariate p (p < 0.1)
logarea	3.98						
logpop		-30.2		-67.79			-63.75
pop18	1.35	1.34	0.84	1.01	0.84	0.8	1.0
pop65	0.29	0.002					-0.01
logdocs			5.14	5.16	5.14	3.53	2.55
logbeds		3.65				1.82	1.95
hsgrad	0.22						-0.28
bagrad	-0.28	-0.33					
logpoverty	18.67	16.8	15.92	15.89	15.92	11.86	15.43
unemp	0.51			1.17			
pcincome	0.002		0.001	-0.003	0.001		-0.003
logtotalinc		31.5		66.91			62.51
region 1	0	0	0	0	0	0	0
region 2	12.32	12.07	12.65	12.85	12.65	11.68	12.08
region 3	25.83	26.83	24.78	27.05	24.78	24.79	25.78
region 4	16.45	17.31	13.85	15.52	13.85	17.03	19.25
logpopden	6.05					1.97	1.52
Test error	16.54	15.07	15.29	17.79	15.29	15.16	14.79

*In region, 1 =northeast, 2 = north central, 3 = south, 4 = west.

*logpopden = log(pop/area), is the log transformation of population density.

Figure 1: Distributions and Loess Curve Fits Before and After Transforming “area”

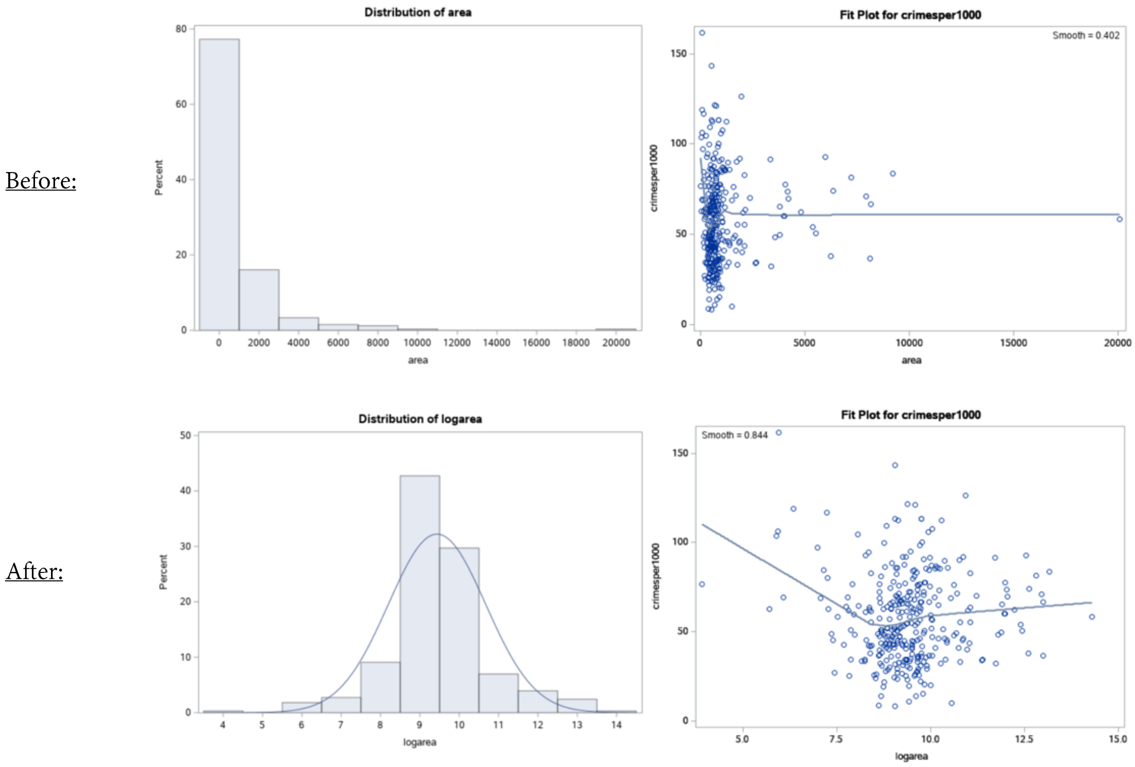


Figure 2: Distributions and Loess Curve Fits Before and After Transforming “pop”

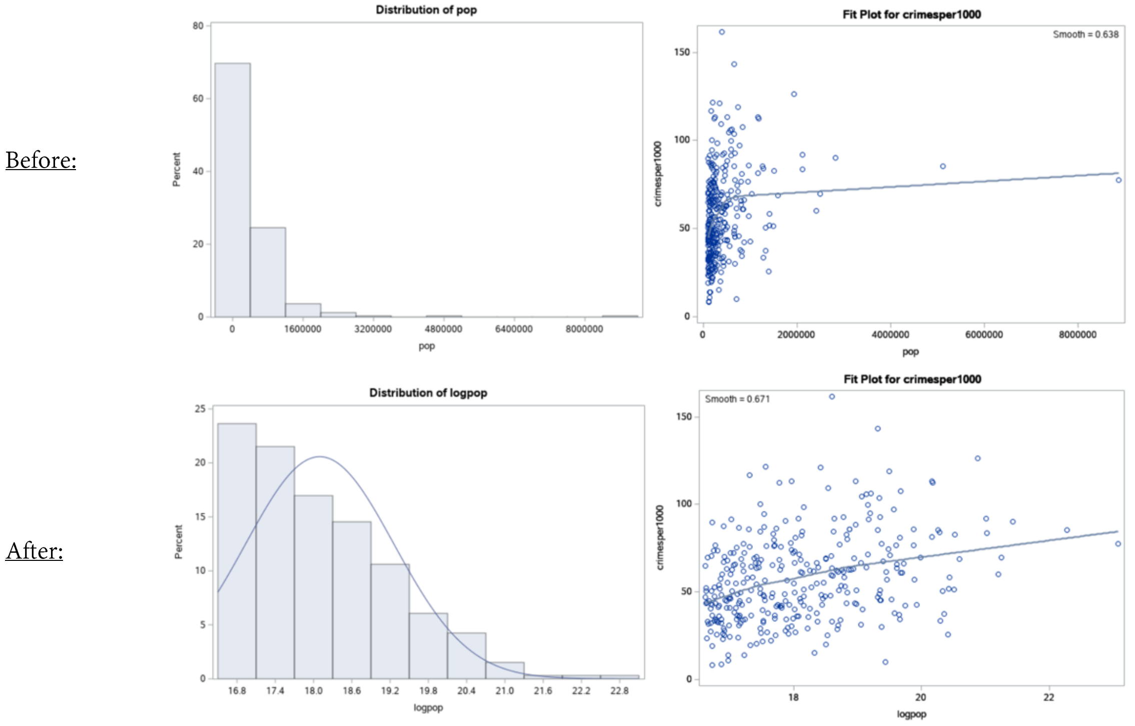


Figure 3: Distributions and Loess Curve Fits Before and After Transforming “docs”

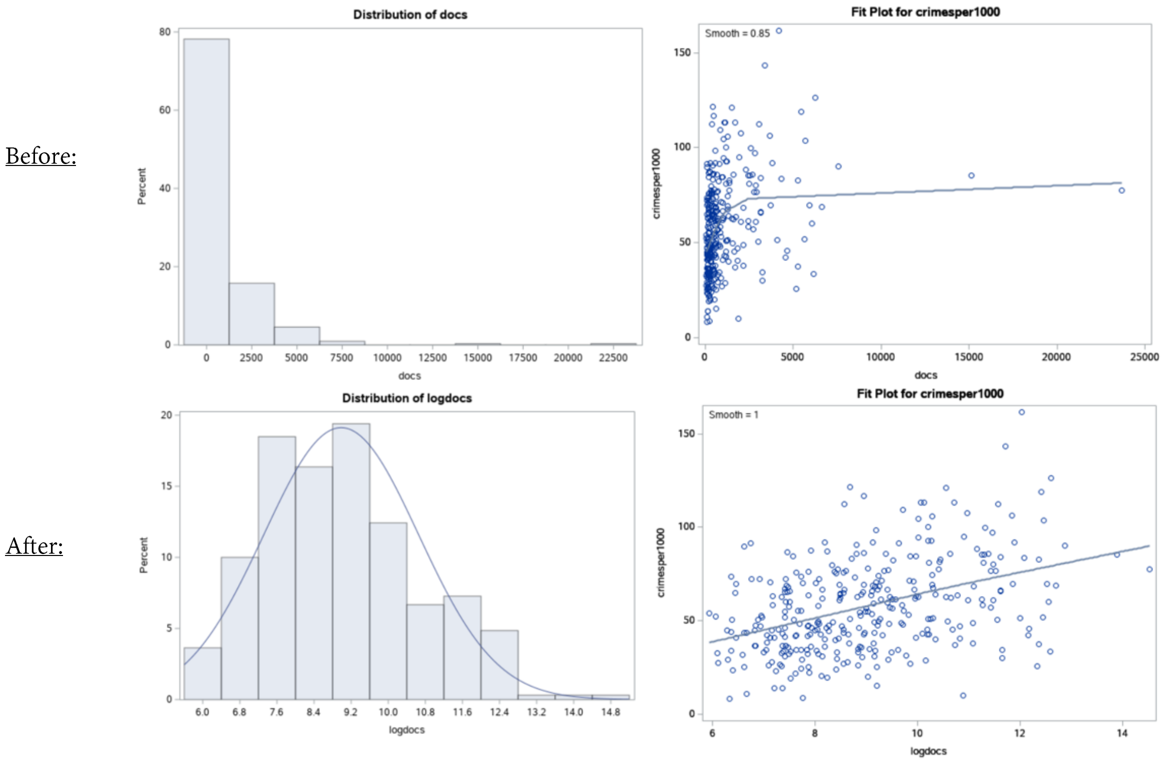


Figure 4: Distributions and Loess Curve Fits Before and After Transforming “beds”

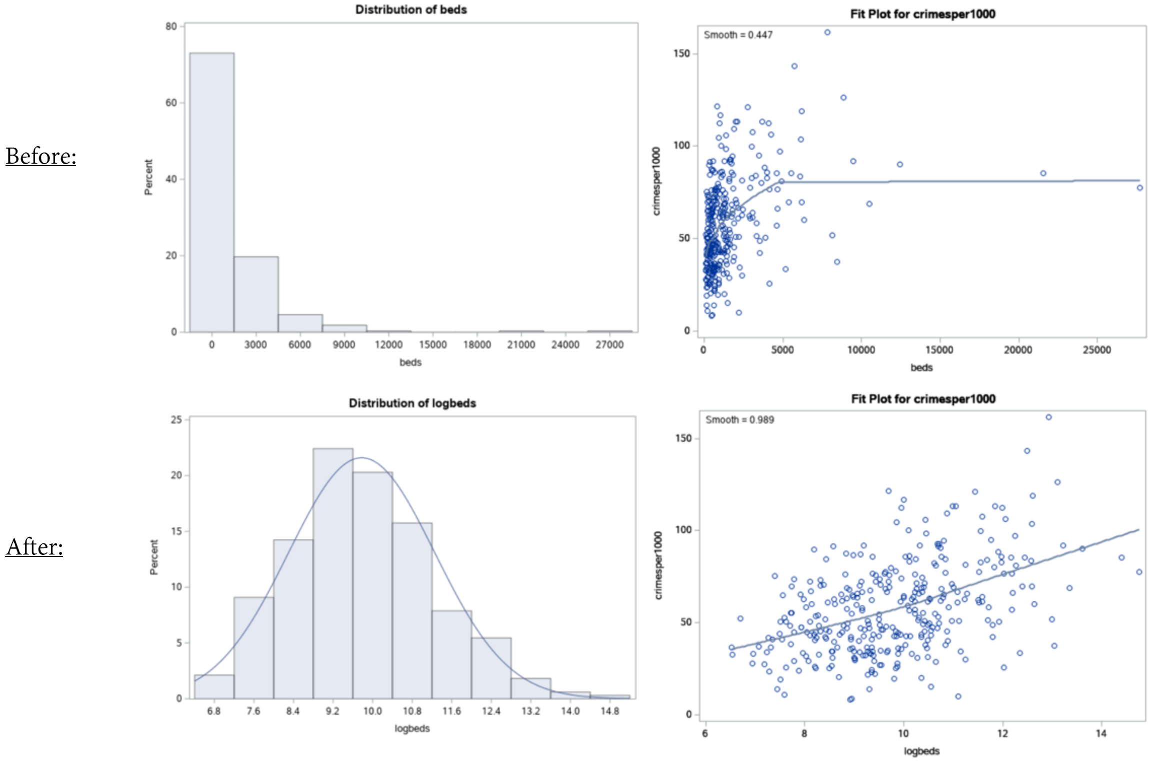
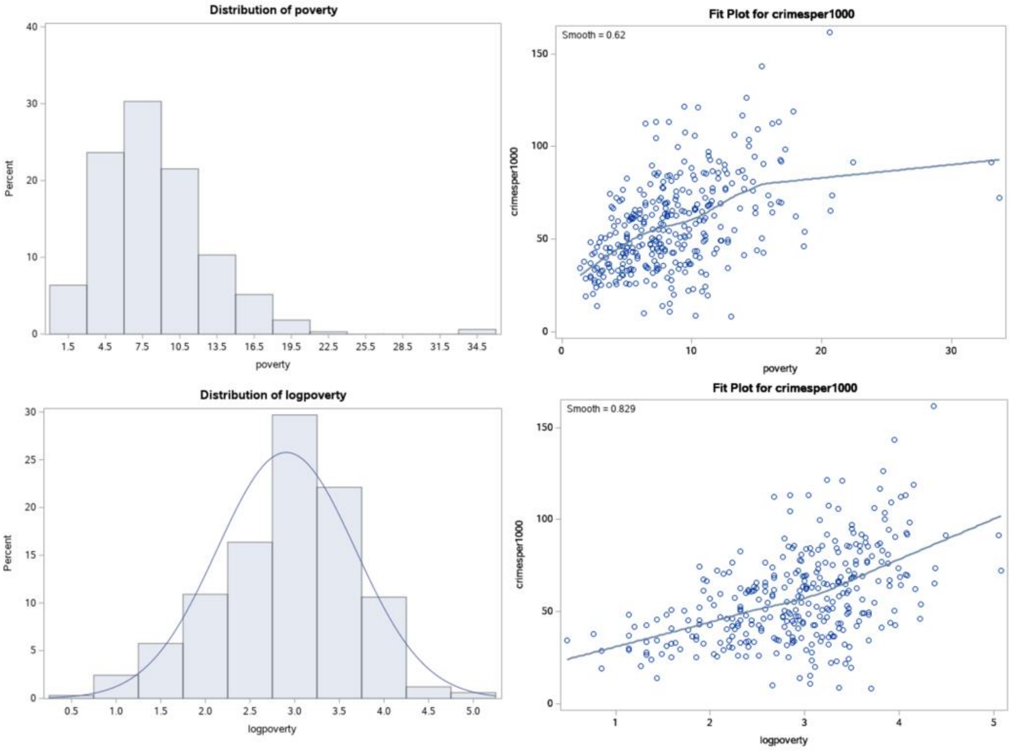


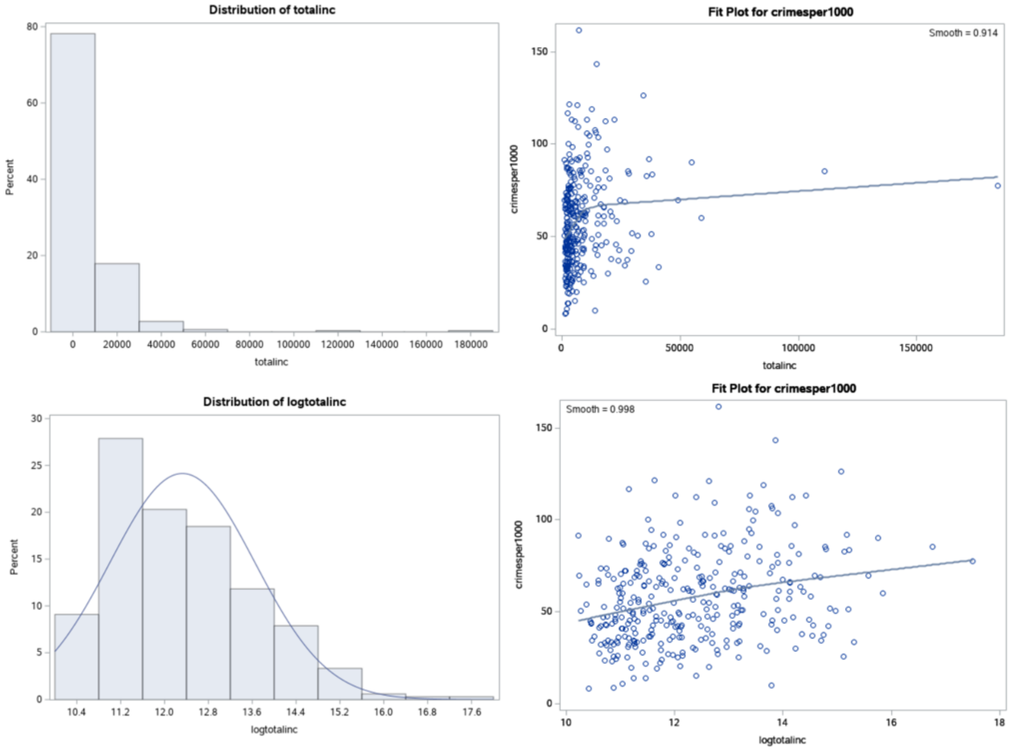
Figure 5: Distributions and Loess Curve Fits Before and After Transforming “poverty”



Before:

After:

Figure 6: Distributions and Loess Curve Fits Before and After Transforming “totalinc”



Before:

After:

Figure 7: Distributions and Loess Curve Fits Before and After Transforming “popden”

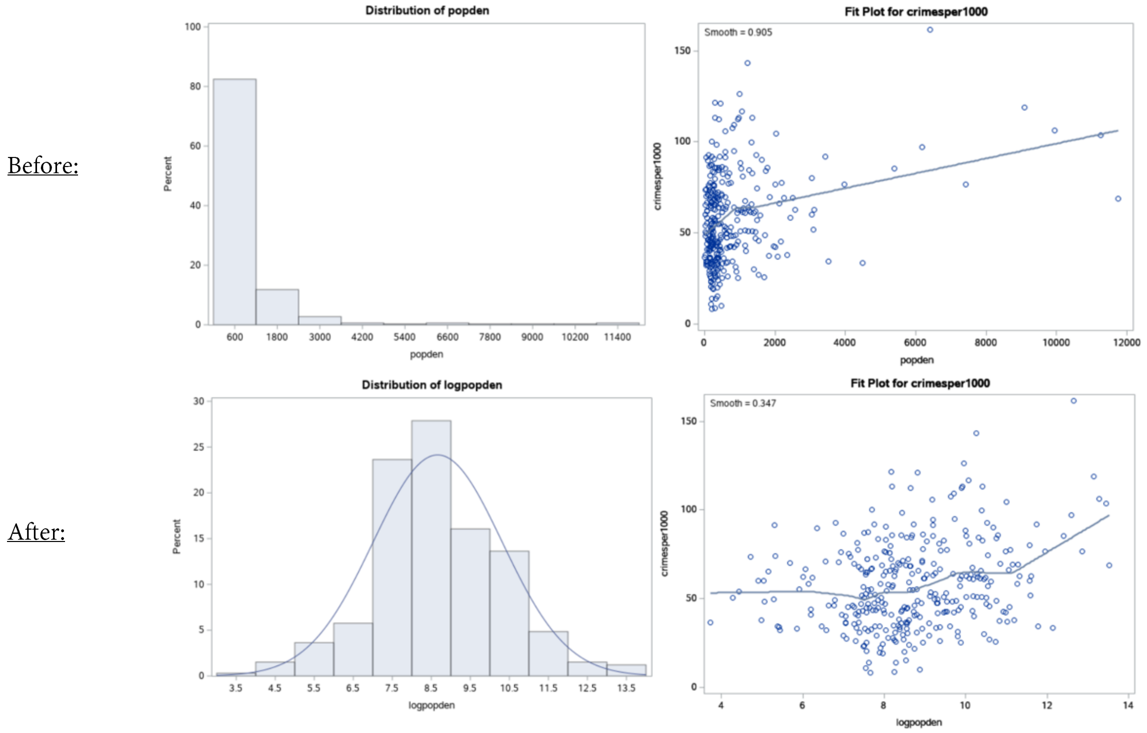
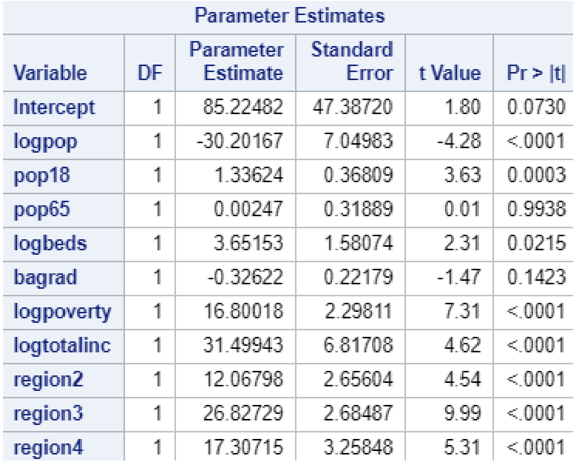
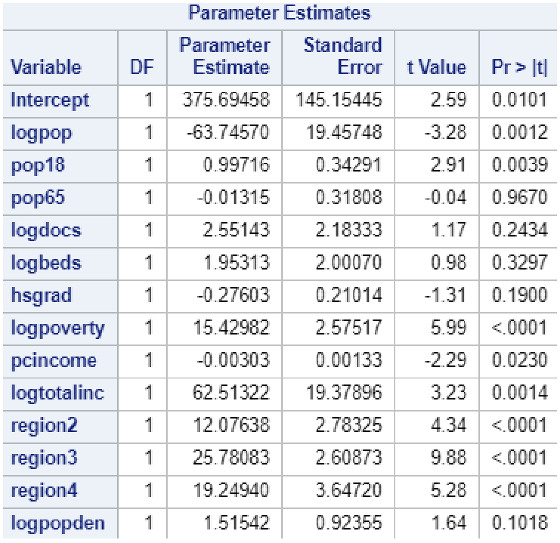


Table 3: Parameter Estimates for Best Subset Selection



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	85.22482	47.38720	1.80	0.0730
logpop	1	-30.20167	7.04983	-4.28	<.0001
pop18	1	1.33624	0.36809	3.63	0.0003
pop65	1	0.00247	0.31889	0.01	0.9938
logbeds	1	3.65153	1.58074	2.31	0.0215
bagrad	1	-0.32622	0.22179	-1.47	0.1423
logpoverty	1	16.80018	2.29811	7.31	<.0001
logtotalinc	1	31.49943	6.81708	4.62	<.0001
region2	1	12.06798	2.65604	4.54	<.0001
region3	1	26.82729	2.68487	9.99	<.0001
region4	1	17.30715	3.25848	5.31	<.0001

Table 4: Parameter Estimates for Bivariate p-value



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	375.69458	145.15445	2.59	0.0101
logpop	1	-63.74570	19.45748	-3.28	0.0012
pop18	1	0.99716	0.34291	2.91	0.0039
pop65	1	-0.01315	0.31808	-0.04	0.9670
logdocs	1	2.55143	2.18333	1.17	0.2434
logbeds	1	1.95313	2.00070	0.98	0.3297
hsgrad	1	-0.27603	0.21014	-1.31	0.1900
logpoverty	1	15.42982	2.57517	5.99	<.0001
pcincome	1	-0.00303	0.00133	-2.29	0.0230
logtotalinc	1	62.51322	19.37896	3.23	0.0014
region2	1	12.07638	2.78325	4.34	<.0001
region3	1	25.78083	2.60873	9.88	<.0001
region4	1	19.24940	3.64720	5.28	<.0001
logpopden	1	1.51542	0.92355	1.64	0.1018