Journal of Civil & Legal Sciences
ISSN: 2169-0170
Research Article Open Access

An Analysis of Predictors of the Total Incarcerated Population in the U.S. using Regression Analysis

Francesca Amato*

Department of applied analytics, Columbia University, New York City, New York

Abstract

Incarceration rates in the United States have been on steady increase since the 1980s. While the U.S. holds about 4% of the world's population, it holds about 22% of the world's incarcerated population. Analyzing the races of these incarcerated populations by state has shown that minority individuals are incarcerated at much higher rates than White individuals. Some may argue that race is a determining factor of whether or not an individual has negative interactions with police, is arrested, and convicted. In this study, regression analysis was used to analyse multiple possible predictor variables of the total incarcerated population in the U.S. to see which, if any are significant predictors.

Keywords: Individuals; Incarceration and U.S. population

Introduction

How did we get here? In the early 1970s, there were less than 300,000 individuals incarcerated in the United States. Today, there are more than 2 million individuals incarcerated with more than 4 million individuals on probation or parole [1]. This exponential growth was the product of the war on drugs that took off in the 1980s. Today, there are more people currently incarcerated for drug related charges than the total amount of incarcerated individuals in 1980. As sentencing has gotten harsher, individuals have been serving longer sentences and the amount of individuals receiving life sentences have been on a steady increase. Currently, 1 in 9 individuals in prison are serving a life sentence [2].

Incarceration rates in the U.S. are not equal amongst different races. While people of color make up only 37% of the U.S. population, they make up about 67% of the U.S. prison population. Not only are people of color more likely to have negative interactions with police, but they are more likely to be arrested, more likely to be convicted, and more likely to be given a harsher sentence. It is likely that 1 in 17 White males will be incarcerated in their lifetime. But for people of color, this ratio is much different. It is likely that 1 in 3 Black men and 1 in 6 Latino men will be incarcerated in their lifetime.

As if imprisonment wasn't worse enough, research has shown that 4 out of 10 individuals that are incarcerated will be re-incarcerated within three years of their release [3]. It is not surprising that the U.S. has the highest incarceration rate in the world. What is surprising is that some of our states having higher incarceration rates than whole nations. There are countries that have higher murder rates than the

U.S. yet somehow, have a smaller incarceration rate. We are supposed to be the land of the free, but incarceration rates say otherwise [4].

Other countries have taken a different approach to crime which is a main reason why their incarceration rates are so much lower. More philosophical and practical approaches have been taken, especially towards individuals who committed non-violent crimes. The U.S. has seemingly become reliant on incarceration to a drastic extent. The worst part about it is that it doesn't seem to affect our crime rate. While crime rates have been declining since the 1990s, studies have found that this is not due to incarceration rates. If incarceration isn't a factor in declining crime rates, it makes you wonder why our incarceration rates are so high.

In a study to analyze race as the predictor of a charge, after controlling all attributes except for race, it was found that race was independently correlated with the severity of the charge [5]. With race being such a big factor in many different aspects of society, this analysis attempts to see if race is a predictor of the total amount of incarcerated individuals in the U.S. alongside some other potential predictors.

Methodology

Using common knowledge, possible predictor variables of the total incarcerated population were determined. These predictors included the total probation population, total parole population, total estimated population, total White population, total Black population, total Hispanic population, median household income, gross domestic product (GDP) in billions, total law enforcement employees, and the violent crime rate. Since people of color are disproportionately affected by incarceration, population totals by race were included. GDP has been used as a measure of the health of an economy. It is the total market value of the finished goods and services produced in an area during a specific time period [6]. It is thought that in a healthy economy, there would be a lower crime rate thus, less individuals incarcerated. As a measure of an individual's/family's financial stability, median household income was also included. It is thought that with a higher median household income comes a greater financial stability thus, individuals in that household are less likely to commit a crime and become incarcerated.

First, variable selection was performed to see which, if any, of the variables are significant predictors of the total incarcerated population. Then the model assumptions were checked and any needed transformations were performed. The model was then checked for outliers and the resulting models were compared. After choosing a model, the regression output was analyzed to determine the fit of the model and the significance of the predictor variables. These processes were performed in Excel and RStudio.

A. Data Sources

For this study, data was collected for each of the 50 states from the

***Corresponding author:** Francesca Amato, Department of applied analytics, Columbia University, New York City, New York, Tel: 05164746326; E-mail: francescaamato932@yahoo.com

Received November 05, 2021; **Accepted** November 19, 2021; **Published** November 26, 2021

Citation: Amato F (2021) An Analysis of Predictors of the Total Incarcerated Population in the U.S. using Regression Analysis. J Civil Legal Sci 10: 301.

year 2019. Multiple sources were used to build the dataset. The total state population [7], total White, Black, and Hispanic state population [8], and median household income by state [9] came from the U.S. Census Bureau. The total state prison population, state probation population, and state prison population, state probation population, and state parole population came from The Sentencing Project [10]. The total amount of law enforcement employees by state came from the Federal Bureau of Investigation (FBI) [11]. The violent crime rate for each state came from Statista [12] which is composed of numbers published by the FBI. Lastly, the GDP for each state came from the Bureau of Economic Analysis (BEA) [13].

3. Variable Selection

Table 1 shows the variables included in this study. By using variable selection, we are able to determine which of these variables should be included in our model. There are two reliable methods used when selecting variables: forward selection and backwards elimination. Both were performed on the set of variables in the table and the results were compared.

Forward Selection

A. When using forward selection, we begin with a model that has no predictor variables. We check the correlation between Y and all of our X's, the predictor variable that has the highest simple correlation with Y gets added into the model. If the regression coefficient for this variable is significant, then it is kept in the model and we search for the next variable. The second variable that we add into the model is the one that has the highest simple correlation with the residuals from the first step. If the regression coefficient for this variable is significant then it is kept and we repeat this process until the added variable has an insignificant coefficient or all of the variables have been added into the model. We terminate the procedure when min(t) < 1 [14]. With a cutoff value of min(t) < 1, we can see from Table 2 that forward selection takes all the predictor variables except for X10.

B. Backwards Elimination

When using backwards elimination, we start with the full model and successively drop one variable at a time. We begin by deleting the variable with the smallest t-Test. If all the t-Tests are significant then we retain all of the variables. We will terminate this procedure once min(|*t*|) > 1. With a cutoff value of min(|*t*|) > 1, we can see from Table 3 that backwards elimination first removes X_{10} it then removes X_4 from the model.

C. Comparing Models

While we would hope for our selection procedures to give the same results, that is not the case here. While both procedures excluded X_{10}, in forward selection X_4 was the first variable to be added but it was removed during backwards elimination. To determine which model to use, the Akaike Information Criterion (AIC) was used. This is a measure that judges the adequacy of a model. The formula for calculating the AIC is as follows:

$$AIC_p = n \ln(SSE_p/n) + 2p$$

where p is the number of variables in the equation, n is the number of observations and SSE_p is the sum of squares error for the equation with p variables. While the AIC for a single model is not very useful, it is a useful tool to rank models. If two models have an AIC that doesn't differ by more than 2, they are equally adequate. For differences larger than 2, the model with the smaller AIC is the one that should be adopted.

In this case we are comparing the model that resulted from forward selection with variables X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, and X_9 to the model that resulted from backwards elimination with the variables X_1, X_2, X_3, X_5, X_6, X_7, X_8, and X_9. The model that resulted from forward selection has an AIC of 900.68 and the model that resulted from backwards elimination has an AIC of 898.98. Since these models only have a difference of 1.7, they can be treated as equally adequate and we can proceed with either. Since X_4 was the first variable to be inserted into the model during forward selection, we will proceed with the model that has X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, and X_9 included.

Variable	Definition
Y	Total Incarcerated Population
X_1	Total Probation Population
X_2	Total Parole Population
X_3	Total Estimated Population
X_4	Total White Population
X_5	Total Black Population
X_6	Total Hispanic Population
X_7	Median Household Income
X_8	GDP in Billions
X_9	Total Law EnforcementEmployees
X_{10}	Violent Crime Rate

Table 1: Variables in Incarceration Data.

Variables in Equation	min(*t*)	p	AIC
X_4X_1	4.12	3	952.78
$X_4X_1X_7$	4.07	4	939.42
$X_4X_1X_7X_6$	3.48	5	929.48
$X_4X_1X_7X_6X_5$	3.02	6	922.06
$X_4X_1X_7X_6X_5X_8$	3.12	7	910.31
$X_4X_1X_7X_6X_5X_8X_2$	2.36	8	906.09
$X_4X_1X_7X_6X_5X_8X_2X_3$	1.60	9	900.76
$X_4X_1X_7X_6X_5X_8X_2X_3X_9$	1.30	10	900.68

Table 2: Variables Selected by Forward Selection.

Variables in Equation	min(*t*)	p	AIC
$X_1X_2X_3X_4X_5X_6X_7X_8X_9$	0.48	10	900.68
$X_1X_2X_3X_5X_6X_7X_8X_9$	1.40	9	898.98

Table 3: Variables Selected by Backwards Elimination .

4. Model Assumptions

The properties of the method of least squares in regression are based on assumptions that we make about our model. These include assumptions about the form of our model, the errors, the predictors, and the observations. We can use multiple plots to analyze whether or not these assumptions hold in our model.

A. Linearity Assumption

The linearity assumption states that the model relating the response variable, Y, to the predictor variables, $X_1, \ldots, X_p$, is assumed to be linear. We can determine if this assumption holds by looking at the scatter plot of the standardized residuals vs. each of the predictor variables. While some of the predictor variables appear to have outliers, our plots show a random scatter for each predictor variable. Thus, we can conclude that linearity holds. These plots can be found in Appendix A.

B. Independent Errors Assumption

This assumptions state that the errors are independent of each other meaning that each observation is independent of each other.

This assumption can be checked by examining the index plot of the standardized residuals. Since our plot has a random scatter, we can conclude that the errors are independent and the assumption holds. The plot can be found in Appendix B.

C. Normality Assumption

This assumption states that the residuals are approximately normally distributed. This can be checked by examining the normal probability plot of the standardized residuals. Since our plot resembles a nearly straight line, we can assume that this assumption holds. The plot can be found in Appendix C.

D. Homoscedasticity Assumption

This assumption states that the errors have the same, but unknown, variance σ^2. This can be checked by examining the scatter plot of the residuals vs. the fitted values. When this plot has a random scatter, the assumption holds. Below is the observed plot for this model.

As opposed to a random scatter, this plot resembles the shape of a cone which is an indication of heteroscedasticity which must be removed.

5. Transforming the Data

In order to remove the heteroscedasticity from our model, we must transform the data. This involves performing a transformation on our response variable in order to stabilize our variance. A common transformation that removes heteroscedasticity is Y which will give a resulting variance

$$W = \sqrt{\_}Y$$

which becomes our response variable. After performing the transformation, the following plot was observed. This plot appears to have no observable pattern with a more random scatter. Thus, we can conclude theteroscedasticity has been removed and the homoscedasticity assumption now holds.

6. Analysis

Now that all of the assumptions hold, we can continue with our analysis. After regressing Y on X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, and X_9, we received the output shown in Table 4. From this output, we can define our model as follows:

$$W = 138.24 + 0.00007X + 0.0005X + 0.00004X - 0.00001X - 0.00000 0X - 0.00001X - 0.001X - 0 \quad X + 0. \quad X \quad .203 \quad 0001$$

To measure the quality of fit of this model we can examine R^2. The closer this value is to unity, the better the fit of the model. We have $R^2 = 0.93$ which tells us that 93% of the variation in the data is accounted for by the model. Thus, we can conclude that there is an excellent fit.

Variable	Coefficient	s.e.	t-Test	p-value
Intercept	138.24	32.84	4.21	0.0001
X1	0.00007	0.0001	0.698	0.49
X2	0.0005	0.0003	1.56	0.13
X3	0.00004	0.00001	2.7	0.01
X4	-0.00001	0.00001	-1.14	0.26
X5	-0.0000001	0.00001	-0.009	0.99
X6	-0.00001	0.000005	-2.2	0.03
X7	-0.001	0.0005	-2.07	0.04
X8	-0.203	0.0685	-2.97	0.005
X9	0.0001	0.00088	0.12	0.9
n = 50	R2 = 0.93	R2 = 0.91	σ̂ =28.7	df = 40

Table 4: Estimated Regression Coefficients.

A. Outliers

To further analyze our model, we can check for any outliers that may exist. Measuring the influence of an observation can tell us if that observation is influential. In this analysis, Cook's distance and the Welsch and Kuh measure, named DFITS, were used to measure the influence of each observation. The formula for Cook's distance is as follows.

$$C_i = \frac{\Sigma n \quad (\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}2(p+1)}, \quad i = 1, 2, \ldots, n$$

When a point is influential, it has a large value of C_i. A common rule is to classify any observations with $C_i > 1$ as influential. We can also use the index plot of these values to determine any possible influential observations. When all C_i values are about the same, then no action needs to be taken. If there are observations that stand out from the rest, they should be flagged and examined.

As opposed to using a strict cutoff value for the observed DFITS values, like Cook's distance, the index plot of the measure can be examined. Observations that stand out from the rest should be flagged and examined [15]. Table 5 shows the Cook's distances and DFITs measures for each observation in our data.

Looking at these measures, we can see that there are a few measures that drastically stand out from the rest. Row 5, which is California, has a Cook's distance of 0.89 and a DFITs measure of -2.969. Row 43, which is Texas, has a Cook's distance of 0.743 and a DFITs measure of 2.74. While these observations have correctly measured data, when they are removed the plots for our assumptions become more satisfactory. Therefore, we will keep these two observations excluded from the data.

Row	C_i	H_i	State	C_i	H_i
1	0.002	-0.149	26	0.0001	-0.035
2	0.000002	0.0045	27	0.00005	-0.02
3	0.005	0.229	28	0.0006	-0.079
4	0.002	-0.149	29	0.0004	0.06
5	0.89	-2.969	30	0.1012	-1.006
6	0.0059	0.241	31	0.0101	-0.318
7	0.0005	-0.07	32	0.013	-0.356
8	0.0001	-0.034	33	0.129	-1.213
9	0.0368	-0.6	34	0.0001	-0.03
10	0.164	1.269	35	0.009	-0.298
11	0.0004	-0.06	36	0.039	0.643
12	0.0003	-0.05	37	0.0026	-0.158
13	0.008	-0.28	38	0.46	-2.165
14	0.002	0.122	39	0.0021	-0.143
15	0.0006	-0.07	40	0.021	-0.462
16	0.0001	0.03	41	0.0002	-0.04
17	0.0157	0.4	42	0.042	0.67
18	0.211	1.636	43	0.743	2.74
19	0.0027	-0.162	44	0.00006	0.024
20	0.015	-0.386	45	0.0012	-0.102
21	0.0045	0.209	46	0.241	1.79
22	0.0156	-0.395	47	0.0101	0.317
23	0.0085	-0.292	48	0.0033	-0.18
24	0.002	-0.139	49	0.0035	0.186
25	0.0006	0.07	50	0.000004	0.007

Table 5: Influence Measures from Fitted Model.

The updated assumption plots can be found in Appendices A, B, C, and D.

While there are still some influential observations in the data after these two outliers are removed, the data from the observations are correct and the variation between states is expected. Thus, we will keep these influential observations in the data. Regressing Y on X_1, ..., X_9 with the two outliers removed gives us the results shown in Table 6.

Variable	Coefficient	s.e.	t-Test	p-value
Intercept	148.29	34.09	4.35	0.0001
X_1	0.00005	0.0001	0.53	0.601
X_2	0.0004	0.0003	1.25	0.22
X_3	0.00007	0.00003	2.68	0.01
X_4	-0.00004	0.00002	-1.77	0.09
X_5	-0.00003	0.00003	-1.13	0.27
X_6	-0.00001	0.00001	-1.31	0.2
X_7	-0.0012	0.0005	-2.36	0.02
X_8	-0.23	0.073	-3.15	0.003
X_9	-0.00006	0.001	-0.06	0.96
n = 48	R^2 = 0.895	R^2 = 0.87	$\hat{\sigma}$ = 25.8	df = 38

Table 6: Estimated Regression Coefficients With Outliers Remove.

Using this output, we can formulate our model which is as follows.

$$W = 148.29+0.00005X_1+0.0004X_2+0.00007_3-0.00004X_4-0.00003X_5-0.00001X_6-0.0012X_7-0.23X_8-0.00006X_9$$

While our R^2 isn't as high as it was with the outliers included, it is still very good with 89.5% of the variation in our data being accounted for by the model.

Results

From this analysis, there are multiple things that the model tells us about the predictor variables we have chosen. The first observation we make is from the variable selection procedures. During both forward selection and backward elimination, X_{10} was not added into the model. It had a t-Test of 0.16 and a p-value of 0.87 which tells us it is not a significant predictor of the response variable which is why it was excluded from the model. This means that the violent crime rate is not a significant predictor of the total incarcerated population.

Our final reduced model excludes X_{10}, uses W as the response variable, and excludes the outlying observations (California and Texas). Looking at this model, we can determine the significance of each of the predictor variables. We will use a critical value of $\alpha = 0.5$, any predictor variables with a p-value less than this will be considered a significant predictor of our response variable. X_1 has a p-value of 0.6, X_4 has a p-value of 0.09, and X_9 has a p-value of 0.96 so we can say these aren't significant predictors. This means that the total probation population, total white population, and total law enforcement employees are not significant predictors of the the total incarcerated population.

Now, the remaining predictor variables can be considered significant predictors of our response variable. X_8 has the smallest p-value of 0.003. This tells us that the GDP (in billions) is the most significant predictor of the total incarcerated population. X_3 and X_7 are the next two most significant predictors with p-values of 0.01 and 0.02, respectively. This tells us that the total estimated population and the median household income are significant predictors of the total incarcerated population.

Lastly, we have X_2, X_5, and X_6 with p-values of 0.22, 0.27, and 0.2, respectively. We can conclude that these variables are also significant predictors of our response variable. This tells us that the total parole population, total Black population, and total Hispanic population are significant predictors of the total incarcerated population.

Conclusion

Based on these results, we can conclude that there are multiple factors that influence the total amount of individuals incarcerated in the United States. Amongst these factors are an individual's race, financial health, and the health of the economy. The GDP and median household income seem to be strong determinants of the total amount of individuals incarcerated.

Prisons receive a very large budget and also add to the economy by providing thousands of jobs. Further research into the economics and finances of prisons would need to be done to see how these factors are influencing the total incarcerated population. Research that focuses on individuals as opposed to states could give more insight as to how a person's finances influences their involvement in crime.

While race is not the most significant predictor of the total incarcerated population, it is still significant. It should further be noted that while the total Black and Hispanic populations are significant predictors of the total incarcerated population, the total White population is not. Considering our earlier statement that people of color are disproportionately affected by incarceration, this discovery is not surprising. Lastly, it should be acknowledged that the violent crime rate is not a significant predictor of the total incarcerated population. This could mean a few things. It could mean that crime does not determine the amount of people that are incarcerated at all but this seems unrealistic. Another possibility is that other types of crime (non-violent) are more significant predictors of the total incarcerated population. This could explain why our incarceration rates are so high if non-violent crimes are committed more than non-violent crimes. Further research on the rates of different types of crime would need to be conducted in order to make any further assumptions.

References

1. Bryan S (2019). Why american prisons owe their cruelty to slavery, The New York Times NY 1-2.
2. Felicity R (2020).Criminal justice facts. The Sentencing Project US: 1-54.
3. Ram S and Alison S (2014). Sentencing and prison practices in Germany and the Netherlands. Federal Sentencing Reporter CUP US 27 1:33-45.
4. P Wagner and W Sawyer (2018). States of incarceration: The global context. Prison Policy Initiative US 1-1.
5. https://mcjalandhar.in/?page_id=827
6. Jennifer M (2012). Implicit bias in the courtroom. UCLA Law Review L A 59:1124–1186.
7. Jason F (2021). Gross domestic product (gdp). Investopedia NY 1-1.
8. State population totals and components of change: 2010-2019.United states census bureau. USA
9. Population and Housing Unit Estimates Datasets: 2010-2019.United states census bureau. USA
10. Median Household Income in the United States: 2010-2019. United States census bureau. USA
11. Public Involvement and Outreach EJTF Report 2020.SEWRPC.USA
12. About Crime in the U.S. (CIUS):2019.FBI:UCR.USA.
13. Gun Control Effectiveness: Do Gun Control Laws Actually Work? : 2021.USA.
14. GDP by State: 2021.BEA.USA.
15. S Chatterjee and A S Hadi (2013). Regression Analysis by Example. 5th edn Wiley US:1-424.