Paper on NLSY97 Data of Incarceration Status

Xiaohan Wu

Spring 2022

1 Introduction

This report describes patterns in monthly incarceration status by race and gender in the year 2002, using NLSY97 publicly available data. Before analyzing the dataset, the raw dataset was mined into a clean dataset with readable variables wanted.

The three variables to be analyzed are `race`, `gender`, and `total_incarcerated`. The variable `race` includes `Black`, `Hispanic`, `Mixed Race`, and `Non-Black or Non-Hispanic`. The variable `total_incarcerated` summarizes each respondent's monthly incarceration status in 2002, representing the respondent's total months of incarceration in 2022.

2 Analysis

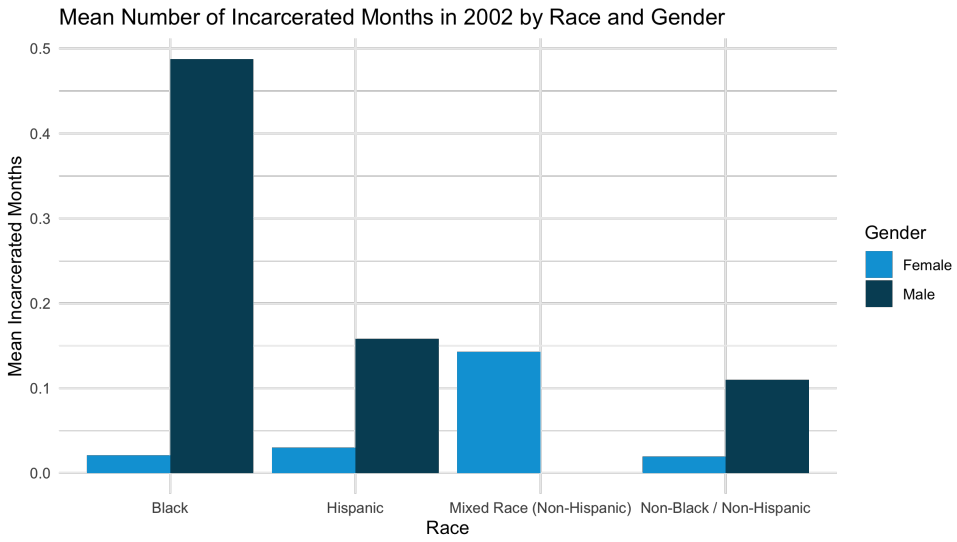

Figure 1: Mean Number of Incarcerated Months in 2002 by Race and Gender

Table 1: Mean Number of incarcerated Months in 2002 by Race and Gender

Gender	Black	Hispanic	Mixed Race Non Hispanic	Non Black Non Hispanic
Female	0.0211268	0.0298013	0.1428571	0.0193192
Male	0.4876712	0.1579509	0.0000000	0.1099476

The Figure 1 above is a barplot showing the mean number of incarcerated months in 2002 by `race` and `gender`, while Table 1 above is showing the detailed values.

For respondents whose variable `race` is `Black`, females' mean incarcerated month number is around 0.02, which is significantly lower than males' mean incarcerated month number (0.5).

For respondents whose variable `race` is `Hispanic`, females' mean incarcerated month number is around 0.03, which is significantly lower than males' mean incarcerated month number (0.16).

For respondents whose variable `race` is `Non-Black or Non-Hispanic`, females' mean incarcerated month number is around 0.02, which is significantly lower than males' mean incarcerated month number (0.11).

For the above three different races, it could be concluded that mainly the males would get much more incarcerations than females in the group of respondents.

However, for respondents whose variable `race` is `Mixed Race Non-Hispanic`, females' mean incarcerated month number is around 0.14, which is significantly higher than males' mean incarcerated month number (0). The result is quite unrealistic since the mean number of incarcerated months is unlikely to be 0 for males in any race category. A possible explanation would be that the gender for `Mixed Race Non-Hispanic` group is mistakenly recorded in the raw dataset. Further investigation and validation would be required and necessary.

Table 2: Regression Output. Omitted category is Black Females.

	Dependent variable:
	Incarcerated Months in 2002
Hispanic	-0.159^{***}
	(0.038)
Mixed Race (Non-Hispanic)	-0.174^{**}
	(0.083)
Non-Black / Non-Hispanic	-0.189^{***}
	(0.035)
Male	0.194^{***}
	(0.022)
Constant	0.155^{***}
	(0.026)
Observations	8,621
R^2	0.015
Adjusted R^2	0.014
Residual Std. Error	1.019 (df = 8616)
F Statistic	32.033^{***} (df = 4; 8616)
Note:	*p<0.1; **p<0.05; ***p<0.01

The table above shows the regression output. The omitted category is Black Females.

The regression result can be shown by the equation below:

$$total\_incarcerated = -0.159Hispanic - 0.174MixedRace - .189Non-BlackorNon-Hispanic + 0.194Male + 0.155$$

The coefficient of `Mixed Race` is sgnificant at 0.05 level, while all the other coefficients are significant at 0.01 level.

However, the regression's R-Squared value is extremely low (around 0,015), indicating the regression may not fit the dataset closely.

3 The Next Steps

First of all, the gender for `Mixed Race Non-Hispanic` group might be mistakenly recorded in the raw dataset. Further investigation and validation would be required and necessary.

The regression's R-Squared value is extremely low (around 0.015), indicating the regression may not fit the dataset closely. Further investigation should be performed.