# My Paper on NLSY97 Data

John Bowman

Sping 2022

## 1 Outline

In this brief study, I tabulated the incarceration status of individuals who responded to the National Longitudinal Survey of Youth in 2002. Using this information, I made a model to predict incarceration status based on racial and gender identity.

The survey collects data on the incarceration status of an individual by month. For the purpose of this study, I generated a binary indicator variable of whether this individual was incarcerated at any point during the year. The survey also collects racial and gender data. The racial categories are Black, Hispanic, Mixed Race Non Hispanic, and Non Black Non Hispanic.

$$Incarcerated = race\beta + gender\beta + \varepsilon$$

The model above predicts the probability of incarceration, based on race and gender. Because each variable is an indicator variable only one race and gender status is assigned to a given observation and is associated with a multiplicative change in probability.

## 2 Analysis

The figure below visualizes the count of individuals who are incarcerated split up by the aforementioned gender and racial groups. It should be noted that the survey includes responses from 8,621 individuals, but there are only a total of 178 displayed in the graph and counted in the subsequent table. Additionally, there are no observations of incarceration for Mixed Race Non Hispanic Males. This will likely cause distortions in the regression since the sample of individuals who are actually incarcerated is simply not large enough for an accurate probability estimate of incarceration.

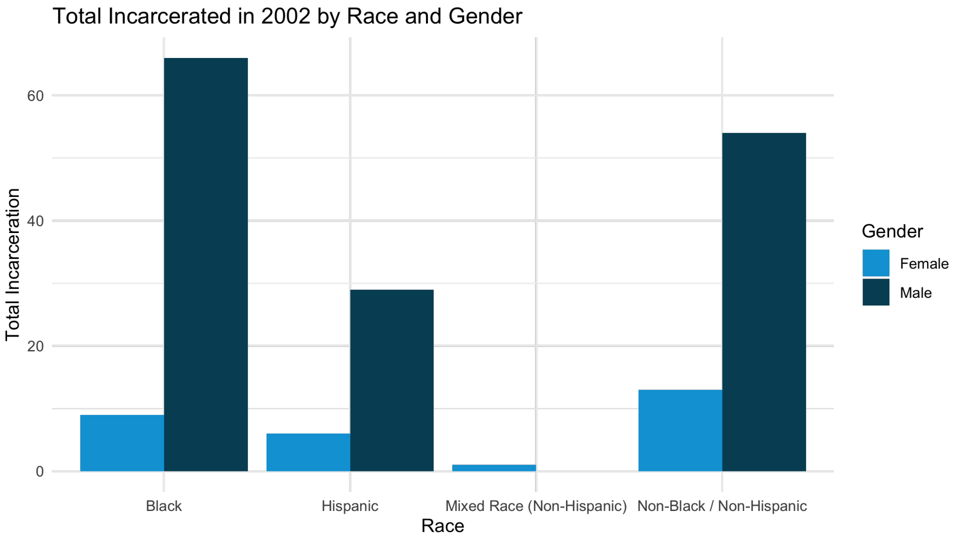

Figure 1: Mean Number of Months Incarcerated in 2002 by Race and Gender (this is the LaTeX caption, not the ggplot title)

Figure 1 does demonstrate some expected qualites. Males are vastly over represented in terms of incarceration status. Ethnic minorites are also overrepresented, compared to the per capita population. The causal relationships for these outcomes are not the subject of this study, but they deserve merit as they could indicate systemic inequites in the justice system. The table below represent the above bar plot numerically.

Table 1: Months incarcerated in 2002 by Race and Gender

| Gender | Black | Hispanic | Mixed Race Non Hispanic | Non Black Non Hispanic |
|---|---|---|---|---|
| Female | 9 | 6 | 1 | 13 |
| Male | 66 | 29 | 0 | 54 |

To interpret these results, take e to the power of the beta value, the probability is multiplied by this exponent. These results show negative probability across the board, with the exception of male. It

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

| | *Dependent variable:* |
|---|---|
| | Incarceration status in 2002 |
| Hispanic | $-0.618^{***}$ |
| | (0.209) |
| Mixed Race (Non-Hispanic) | $-1.021$ |
| | (1.036) |
| Non-Black / Non-Hispanic | $-0.866^{***}$ |
| | (0.172) |
| Male | $1.660^{***}$ |
| | (0.205) |
| Constant | $-4.470^{***}$ |
| | (0.197) |
| Observations | 8,621 |
| Log Likelihood | $-810.209$ |
| Akaike Inf. Crit. | 1,630.419 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

is likely that the negative probability are due to a overwhelming number of individuals in the data set who were not incarcerated. It is likely that this data set is not large enough to properly predict these probability.