# Statistical Discrimination in Law Enforcement

James Freeland

2024-05-02

## I. Introduction

Statistical discrimination within law enforcement is the use of empirical data and statistical algorithms to predict the likelihood of criminal activity based on group characteristics. It is used to increase the objectivity and efficiency of policing practices. However, statistical discrimination raises serious ethical and accuracy concerns that are exacerbated by model bias and imperfect information. Historically, males, specifically of African or aboriginal descent and those who are financially insecure or mentally ill tend to have higher rates of conviction than their counterparts, holding all else equal. Statistical discrimination became a legal precedent in the 1993 case, Whren v. United States, when two black men were stopped while driving and subsequently arrested for possession of cocaine despite legal provocation for the stop in the first place.

In pursuit of understanding the influence of group characteristics on arrests and convictions, Ambrose Leung et al. performed a study in 2005 that tested survey data of Montréal youths self-reporting crime and various other characteristics in order to determine if police overuse personal information on individuals when investigating criminal activity. After data cleaning, the survey yielded 639 individuals who, at age, lived in lower-income regions of Montréal. Due to the relatively isolated area of study, those sampled were fairly homogenous in immutable factors such as race and sex, leading to the omission of such variables within their models. Variables that were used include schooling, criminal and delinquent associates, parental characteristics, family type, employment, and gang membership.

The researchers regressed three logit models on group characteristics to predict the probability of showing up in court, as a proxy for criminality. The initial model was regressed upon self-reported delinquencies, the second on individual and family characteristics, and the third is a combination of the two. In their third regression, when controlling for delinquent behavior, schooling, and other individual characteristics still had significant effects on the likelihood of making a court appearance. This is counter to the pursuit of pure, informed statistical discrimination as one would expect group characteristics to become futile within the model. This insinuates that law enforcement overuses specific familial and individual characteristics of criminal behavior, underscoring a normative concern about the disproportionate use of group characteristics when pursuing and apprehending individuals.

## II. Analysis of Methods

In an attempt to substantiate these findings, I tried to create a similar series of models on the 1997 National Longitudinal Survey of Youth (NLSY97). The NLSY97 followed a group of young individuals for many years as they transitioned into adulthood, gathering information on self-reported delinquent behavior, individual and family characteristics, and legal engagements. There were initially 8,900 respondents of diverse backgrounds from around the United States. The data contained intratemporal responses for most features that were recorded for each year of the individual's life. For my study, I restricted intratemporal indicator regressors to ages 15 through 20 for respondent consistency.

In my pursuit, I created three vectors of characteristic features. The first is the delinquent vector consisting of self-reports of marijuana and alcohol use, theft over $50, assaultive behavior, the sale of illegal drugs, destruction of property, and gang membership across the ages of 15 to 20. The second vector is individual characteristics including education enrollment, highest grade completed, ASVAB score, and weeks worked in a given year. The third is family characteristics including family structure, employment status of their mother and father, and parental education.

Three logit regressions are modeled from these vectors: the first is regressed solely upon the delinquency vector; the second upon individual and familial characteristics; and the third from a combination of the two. A logit model is the best fit to model this data due to its ability to handle binary outcomes. Due to the nature of many regressors being intratemporal indicators, there was a serious concern about multicollinearity amongst the variables. To reconcile this I computed a Variance Inflation Factor (VIF) calculation on the regressed models. This revealed that the intratemporal series of the highest grade completed (HGC) variables were highly collinear, leading to the decision to only include the HGC at the age of twenty. All other intratemporal variable series proved to not be highly collinear with GVIF scores less than or equal to 3, which was determined to be sufficient for this analysis.

Table 1: Delinquency Model Summary

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -3.264 | 0.399 | -8.173 | 0.000 |
| DESTROY_CY15(1) Yes | -0.703 | 0.513 | -1.371 | 0.170 |
| DESTROY_CY16(1) Yes | -0.458 | 0.536 | -0.854 | 0.393 |
| DESTROY_CY17(1) Yes | 0.157 | 0.556 | 0.283 | 0.777 |
| DESTROY_CY18(1) Yes | 0.731 | 0.654 | 1.117 | 0.264 |
| DESTROY_CY19(1) Yes | -0.517 | 0.747 | -0.693 | 0.488 |
| DESTROY_CY20(1) Yes | 0.576 | 0.651 | 0.886 | 0.376 |
| STEAL_OVER_50_CY15(1) Yes | 3.354 | 0.958 | 3.502 | 0.000 |
| STEAL_OVER_50_CY16(1) Yes | 1.160 | 0.687 | 1.689 | 0.091 |
| STEAL_OVER_50_CY17(1) Yes | 1.017 | 0.876 | 1.161 | 0.245 |
| STEAL_OVER_50_CY18(1) Yes | 0.954 | 0.953 | 1.002 | 0.317 |
| STEAL_OVER_50_CY19(1) Yes | 0.769 | 0.974 | 0.790 | 0.429 |
| STEAL_OVER_50_CY20(1) Yes | 1.074 | 1.071 | 1.003 | 0.316 |
| ASSAULT_CY15(1) Yes | 0.818 | 0.493 | 1.660 | 0.097 |
| ASSAULT_CY16(1) Yes | 1.071 | 0.482 | 2.224 | 0.026 |
| ASSAULT_CY17(1) Yes | 0.592 | 0.544 | 1.089 | 0.276 |
| ASSAULT_CY18(1) Yes | 0.726 | 0.612 | 1.187 | 0.235 |
| ASSAULT_CY19(1) Yes | 0.009 | 0.647 | 0.013 | 0.989 |
| ASSAULT_CY20(1) Yes | -0.576 | 0.781 | -0.737 | 0.461 |
| SELL_DRUGS_CY15(1) Yes | -0.152 | 0.755 | -0.202 | 0.840 |
| SELL_DRUGS_CY16(1) Yes | -0.469 | 0.670 | -0.699 | 0.484 |
| SELL_DRUGS_CY17(1) Yes | -0.109 | 0.648 | -0.167 | 0.867 |
| SELL_DRUGS_CY18(1) Yes | -0.069 | 0.660 | -0.105 | 0.917 |
| SELL_DRUGS_CY19(1) Yes | -0.587 | 0.698 | -0.841 | 0.400 |
| SELL_DRUGS_CY20(1) Yes | 0.484 | 0.685 | 0.707 | 0.480 |
| IN_A_GANG_CY15(1) Yes | 0.941 | 1.739 | 0.541 | 0.589 |
| IN_A_GANG_CY16(1) Yes | 0.764 | 1.483 | 0.515 | 0.606 |
| IN_A_GANG_CY17(1) Yes | -3.378 | 2.747 | -1.230 | 0.219 |
| IN_A_GANG_CY18(1) Yes | -0.190 | 1.614 | -0.118 | 0.906 |
| IN_A_GANG_CY19(1) Yes | -1.964 | 2.020 | -0.972 | 0.331 |
| IN_A_GANG_CY20(1) Yes | 3.256 | 2.691 | 1.210 | 0.226 |
| MARIJUANA_CY15(1) Yes | 0.551 | 0.429 | 1.282 | 0.200 |
| MARIJUANA_CY16(1) Yes | 1.484 | 0.422 | 3.514 | 0.000 |
| MARIJUANA_CY17(1) Yes | -0.066 | 0.444 | -0.148 | 0.883 |
| MARIJUANA_CY18(1) Yes | 0.352 | 0.424 | 0.830 | 0.407 |
| MARIJUANA_CY19(1) Yes | 0.424 | 0.419 | 1.012 | 0.311 |
| MARIJUANA_CY20(1) Yes | 0.168 | 0.422 | 0.399 | 0.690 |
| ALCOHOL_CY15(1) Yes | -0.001 | 0.388 | -0.003 | 0.998 |
| ALCOHOL_CY16(1) Yes | -0.415 | 0.409 | -1.015 | 0.310 |
| ALCOHOL_CY17(1) Yes | 0.073 | 0.427 | 0.171 | 0.864 |
| ALCOHOL_CY18(1) Yes | -0.044 | 0.426 | -0.102 | 0.918 |
| ALCOHOL_CY19(1) Yes | 0.898 | 0.490 | 1.834 | 0.067 |
| ALCOHOL_CY20(1) Yes | -0.326 | 0.432 | -0.753 | 0.451 |

Within the first regression, solely on delinquent behaviors, the variables deemed statistically significant with positive influence in predicting the likelihood of being arrested were stealing over $50 at both ages 15 and 16, assaultive crimes at ages 15 and 16, marijuana consumption at age 16, and alcohol consumption at age 19. This seems to suggest that participating in delinquent activities at a younger age increases the likelihood of being arrested. Alcohol consumption at age 19 being significant was slightly surprising but could suggest behaviors perceived as less severe in earlier years become more scrutinized by law enforcement as individuals approach legal adulthood.

Table 2: Familial Model Summary

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -10.039 | 1455.398 | -0.007 | 0.994 |
| FMSTRC97(2) One biological- and one step-parent | -0.047 | 0.333 | -0.142 | 0.887 |
| FMSTRC97(5) Other family type | -13.351 | 1455.398 | -0.009 | 0.993 |
| MOMEMP97(1) Yes, employed | 0.076 | 0.329 | 0.230 | 0.818 |
| MOMEMP97(2) Never employed | 0.340 | 0.775 | 0.438 | 0.661 |
| DADEMP97(1) Yes, employed | -0.421 | 0.399 | -1.056 | 0.291 |
| DADEMP97(2) Never employed | -13.427 | 1018.712 | -0.013 | 0.989 |
| RESPARED(2) Graduated from hs | -0.426 | 0.438 | -0.972 | 0.331 |
| RESPARED(3) Some college | 0.081 | 0.456 | 0.179 | 0.858 |
| RESPARED(4) College degree or higher | -0.131 | 0.488 | -0.268 | 0.789 |
| ENROLL_ED_CY15(1) Yes | 14.394 | 1455.398 | 0.010 | 0.992 |
| ENROLL_ED_CY17(1) Yes | 0.349 | 0.959 | 0.364 | 0.716 |
| ENROLL_ED_CY18(1) Yes | -0.461 | 0.547 | -0.844 | 0.399 |
| ENROLL_ED_CY19(1) Yes | -0.561 | 0.332 | -1.687 | 0.092 |
| ENROLL_ED_CY20(1) Yes | -0.291 | 0.338 | -0.863 | 0.388 |
| WKS_WK_CY15 | -0.006 | 0.012 | -0.456 | 0.648 |
| WKS_WK_CY16 | -0.008 | 0.011 | -0.731 | 0.465 |
| WKS_WK_CY17 | 0.010 | 0.009 | 1.053 | 0.292 |
| WKS_WK_CY18 | 0.001 | 0.010 | 0.116 | 0.907 |
| WKS_WK_CY19 | -0.010 | 0.011 | -0.940 | 0.347 |
| WKS_WK_CY20 | -0.002 | 0.009 | -0.206 | 0.837 |
| HGC_CY20 | -0.411 | 0.130 | -3.164 | 0.002 |
| ASVAB_SCORE | 0.007 | 0.006 | 1.159 | 0.246 |

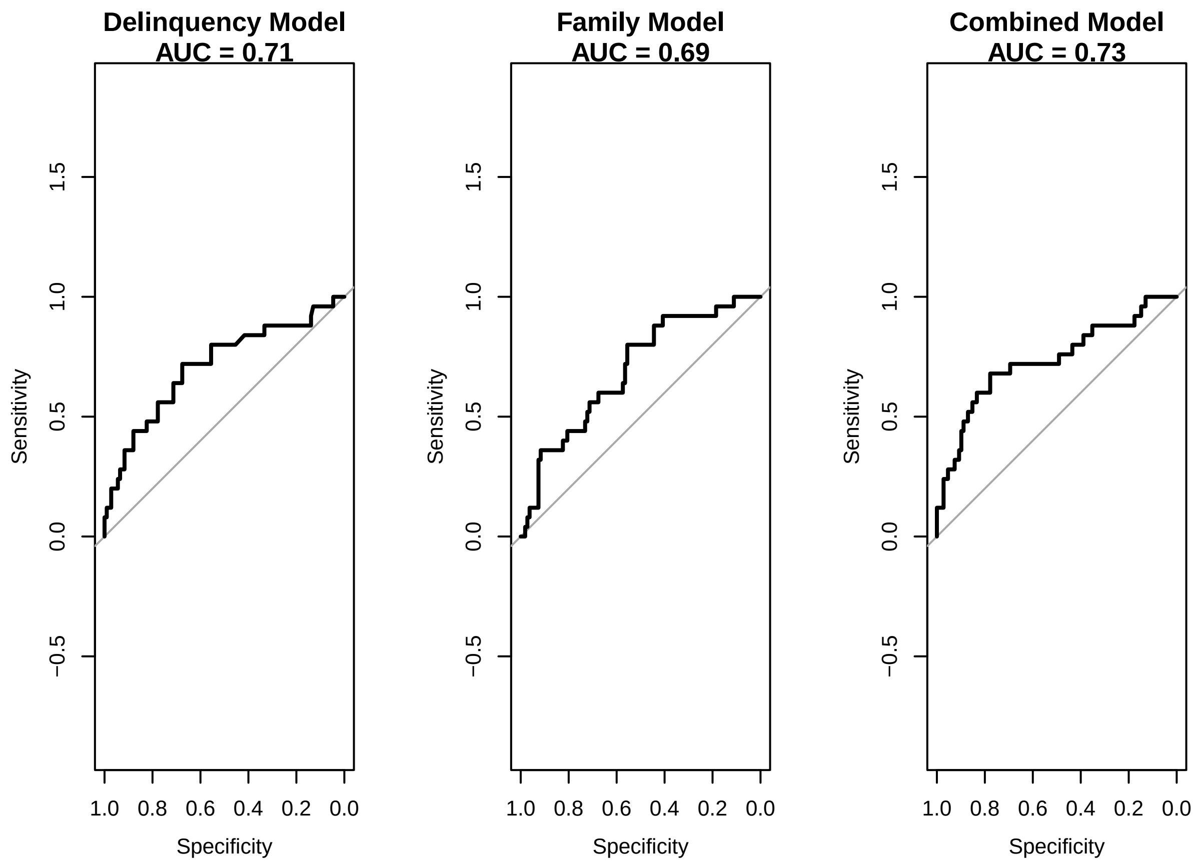
In the second regression on individual and familial attributes, enrollment in education at age 19 was significant and decreased the likelihood of an individual being arrested. The highest grade completed by age 20 was the most significant indicator and also suggested that the more education an individual received, the less likely they are to be arrested.

In the combined regression stealing over $50 at the age of 15 remained highly significant. However, such theft at age 19 became slightly less significant as opposed to age 16. Assaultive behaviors were significant at age 16 and marijuana consumption was significant at ages 16 and 19. Once again, alcohol consumption at age 19 is a significant predictor. These trends seem to corroborate that beginning to engage in delinquent behavior at younger ages increases the likelihood of being arrested. Additionally, the significance at age 19 may suggest maintaining delinquent behavior into early adulthood also increases the likelihood of arrest. The only individual characteristic that maintained significance is education enrollment at age 19 decreasing the log odds of being arrested. This corroborates Leung's finding that police may engage with students differently than non-students when policing, specifically with those at age 19 who are likely to be early in their higher education careers.

It should be noted all three models were fairly accurate, predicting correctly around 80% of the time; additionally, I fitted receiver operating characteristic curves (ROC) to the models, with each model capturing about 70% of the area under the curve; and when computing a confusion matrix, the models tended to commit false negative predictions at higher frequencies than false positive. This pattern of results underscores the robustness of the models in identifying non-arrest cases accurately, while also highlighting a potential area for improvement in reducing false negatives.

Table 3: Combined Model Summary

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -12.427 | 1455.399 | -0.009 | 0.993 |
| DESTROY_CY15(1) Yes | -0.820 | 0.576 | -1.422 | 0.155 |
| DESTROY_CY16(1) Yes | -0.221 | 0.592 | -0.374 | 0.708 |
| DESTROY_CY17(1) Yes | 0.294 | 0.613 | 0.481 | 0.631 |
| DESTROY_CY18(1) Yes | 0.985 | 0.688 | 1.432 | 0.152 |
| DESTROY_CY19(1) Yes | -0.332 | 0.787 | -0.422 | 0.673 |
| DESTROY_CY20(1) Yes | 0.366 | 0.726 | 0.504 | 0.614 |
| STEAL_OVER_50_CY15(1) Yes | 3.440 | 1.024 | 3.358 | 0.001 |
| STEAL_OVER_50_CY16(1) Yes | 1.094 | 0.771 | 1.419 | 0.156 |
| STEAL_OVER_50_CY17(1) Yes | 0.452 | 0.969 | 0.467 | 0.641 |
| STEAL_OVER_50_CY18(1) Yes | 1.072 | 1.139 | 0.941 | 0.347 |
| STEAL_OVER_50_CY19(1) Yes | 1.745 | 1.028 | 1.698 | 0.090 |
| STEAL_OVER_50_CY20(1) Yes | 0.704 | 1.227 | 0.574 | 0.566 |
| ASSAULT_CY15(1) Yes | 0.302 | 0.556 | 0.543 | 0.587 |
| ASSAULT_CY16(1) Yes | 1.141 | 0.528 | 2.163 | 0.031 |
| ASSAULT_CY17(1) Yes | 0.523 | 0.588 | 0.890 | 0.374 |
| ASSAULT_CY18(1) Yes | 0.429 | 0.681 | 0.630 | 0.528 |
| ASSAULT_CY19(1) Yes | -0.121 | 0.751 | -0.162 | 0.872 |
| ASSAULT_CY20(1) Yes | -0.350 | 0.859 | -0.408 | 0.683 |
| SELL_DRUGS_CY15(1) Yes | 0.120 | 0.828 | 0.145 | 0.885 |
| SELL_DRUGS_CY16(1) Yes | -0.189 | 0.763 | -0.248 | 0.804 |
| SELL_DRUGS_CY17(1) Yes | -0.061 | 0.710 | -0.086 | 0.931 |
| SELL_DRUGS_CY18(1) Yes | -0.051 | 0.725 | -0.071 | 0.944 |
| SELL_DRUGS_CY19(1) Yes | -0.570 | 0.752 | -0.758 | 0.448 |
| SELL_DRUGS_CY20(1) Yes | 0.022 | 0.771 | 0.029 | 0.977 |
| IN_A_GANG_CY15(1) Yes | 0.866 | 2.102 | 0.412 | 0.680 |
| IN_A_GANG_CY16(1) Yes | -0.020 | 1.762 | -0.012 | 0.991 |
| IN_A_GANG_CY17(1) Yes | -3.284 | 2.295 | -1.431 | 0.152 |
| IN_A_GANG_CY18(1) Yes | 0.161 | 1.585 | 0.101 | 0.919 |
| IN_A_GANG_CY19(1) Yes | -2.534 | 1.975 | -1.283 | 0.199 |
| IN_A_GANG_CY20(1) Yes | 4.252 | 3.524 | 1.206 | 0.228 |
| MARIJUANA_CY15(1) Yes | 0.507 | 0.472 | 1.075 | 0.282 |
| MARIJUANA_CY16(1) Yes | 1.180 | 0.455 | 2.591 | 0.010 |
| MARIJUANA_CY17(1) Yes | 0.091 | 0.475 | 0.192 | 0.848 |
| MARIJUANA_CY18(1) Yes | 0.150 | 0.467 | 0.321 | 0.748 |
| MARIJUANA_CY19(1) Yes | 0.913 | 0.460 | 1.983 | 0.047 |
| MARIJUANA_CY20(1) Yes | 0.343 | 0.463 | 0.740 | 0.459 |
| ALCOHOL_CY15(1) Yes | -0.058 | 0.412 | -0.140 | 0.889 |
| ALCOHOL_CY16(1) Yes | -0.494 | 0.448 | -1.103 | 0.270 |
| ALCOHOL_CY17(1) Yes | 0.174 | 0.460 | 0.377 | 0.706 |
| ALCOHOL_CY18(1) Yes | -0.188 | 0.460 | -0.410 | 0.682 |
| ALCOHOL_CY19(1) Yes | 0.888 | 0.513 | 1.730 | 0.084 |
| ALCOHOL_CY20(1) Yes | 0.082 | 0.479 | 0.171 | 0.864 |
| FMSTRC97(2) One biological- and one step-parent | 0.149 | 0.434 | 0.343 | 0.732 |
| FMSTRC97(5) Other family type | -12.251 | 1455.398 | -0.008 | 0.993 |
| MOMEMP97(1) Yes, employed | -0.186 | 0.412 | -0.450 | 0.652 |
| MOMEMP97(2) Never employed | 0.615 | 1.053 | 0.584 | 0.559 |
| DADEMP97(1) Yes, employed | -0.038 | 0.548 | -0.070 | 0.944 |
| DADEMP97(2) Never employed | -11.819 | 1018.407 | -0.012 | 0.991 |
| RESPARED(2) Graduated from hs | -0.753 | 0.544 | -1.382 | 0.167 |
| RESPARED(3) Some college | -0.318 | 0.567 | -0.560 | 0.576 |
| RESPARED(4) College degree or higher | -0.674 | 0.599 | -1.126 | 0.260 |
| ENROLL_ED_CY15(1) Yes | 13.035 | 1455.398 | 0.009 | 0.993 |
| ENROLL_ED_CY17(1) Yes | -0.149 | 1.217 | -0.122 | 0.903 |
| ENROLL_ED_CY18(1) Yes | -0.420 | 0.717 | -0.585 | 0.558 |
| ENROLL_ED_CY19(1) Yes | -0.777 | 0.429 | -1.809 | 0.071 |
| ENROLL_ED_CY20(1) Yes | -0.006 | 0.446 | -0.015 | 0.988 |
| WKS_WK_CY15 | 0.008 | 0.015 | 0.509 | 0.611 |
| WKS_WK_CY16 | -0.021 | 0.014 | -1.485 | 0.137 |
| WKS_WK_CY17 | 0.010 | 0.012 | 0.843 | 0.399 |
| WKS_WK_CY18 | 0.017 | 0.013 | 1.331 | 0.183 |
| WKS_WK_CY19 | -0.022 | 0.014 | -1.643 | 0.100 |
| WKS_WK_CY20 | -0.013 | 0.012 | -1.069 | 0.285 |
| HGC_CY20 | -0.138 | 0.172 | -0.802 | 0.422 |
| ASVAB_SCORE | -0.003 | 0.007 | -0.459 | 0.646 |




III. Analysis of Normative Consideration

In evaluating the normative considerations surrounding the use of statistical methods in law enforcement, there is a trend of the overuse of specific familial and individual characteristics as indications of criminal activity. It is important to employ philosophical frameworks such as utilitarianism, deontology, and the harm principle to examine the ethical facets of using statistical discrimination.

From a utilitarian viewpoint, the use of statistical methods in determining the likelihood of criminal behavior can increase the efficiency of resource allocation when policing. If executed properly this optimizes the deployment of limited resources, and could potentially reduce overall crime rates effectively, minimizing harm within communities. However, this gained efficiency has to be weighed against the potential harm caused by inaccuracies and biases within these statistical models. If the predictive models are created based on biased data, this will cause continually greater bias as the models reinforce and perpetuate existing prejudices, leading to a cycle where specific demographics are disproportionately targeted, thereby exacerbating social inequalities. Since it is impossible for officers to work with perfect information this guarantees policing, as a function of group characteristics, inadvertently prioritizes efficiency over fairness. While this seems like a net good in the short run, this can lead to long-run social harm by undermining trust and equity within communities.

The harm principle which imposes that agency should only be restricted to prevent harm to others, would similarly suggest that the use of statistical discrimination could be justified if used to prevent harm, but it becomes unjust when it inevitably leads to harm through bias and perpetuation of stereotypes.

From a deontological focus, the concern is the morality of actions and not just their outcomes. The overuse of demographic characteristics can be considered inherently unethical, regardless of benefit as it treats individuals as a means to an end. In other words, infringing on the liberties of specific individuals for societal benefit is amoral, and when you generalize this action to an entire population it becomes clear such practices could institutionalize discrimination, perpetuating injustice and eroding the moral fabric of law enforcement by violating the inherent rights of individuals. In conclusion, the considerations surrounding the use of statistical discrimination methods in policing should and do extend beyond societal efficiency. Misclassifying individuals as likely offenders based on imperfect analysis of group characteristics harms those individuals without cause and erodes the moral authority and legitimacy of law enforcement. Therefore, even though statistical methods, including machine learning, can give valuable insights, they must be implemented with ethical oversight and review so as to not contribute to societal harm.

## IV. Conclusion

In this paper, we closely examined the statistical and ethical implications of using statistical discrimination in law enforcement. Leung et al. found that multiple individual and group characteristics were overused in the policing of Montréal youths, which can lead to disproportionate and discriminatory enforcement that targets specific demographics. In my recreation of this study across the entire United States, I have found similar, albeit weaker, corroborative evidence that suggests that membership in education beyond the years of high school, specifically at the age of 19, plays a significant role in determining the likelihood of being arrested. Based on these findings, it is recommended that law enforcement agencies institute stricter ethical oversight on internalized and statistical bias to foster fairness, prevent the perpetuation of stereotypes, and stop the erosion and perceived authority of policing bodies. Future studies should further explore and substantiate the long-term effects of statistical discrimination and how it influences public trust and the behaviors of law enforcement and the communities they serve.