Are Police Biased? An NLP Approach

Anonymous Author(s)
Affiliation
Address
email

Abstract

Researchers have traditionally run regressions on numerical and categorical data to detect police bias and inform decisions about criminal justice. This approach can only control for a limited set of simple features, leaving significant unexplained variation and raising concerns of omitted variable bias. Using a novel dataset of text from more than a million police stops, we propose a new method applying large language models (LLMs) to incorporate textual data into regression analysis of stop outcomes. Our LLM-boosted approach has considerably more explanatory power than traditional methods and substantially changes inferences about police bias on characteristics like gender, race, and ethnicity. It also allows us to investigate what features of police reports best predict stops and how officers differ in their conduct of stops. Incorporating textual data ultimately permits more accurate and more detailed inferences on criminal justice data.

1 Introduction

Our criminal justice system relies heavily on prediction, from juvenile crime prevention to police positioning and recidivism assessment. Traditionally, these predictions use numerical or categorically coded data. However, the stakes of these predictions are immense, impacting billions of dollars and countless lives.

This paper contributes to criminal justice prediction research by employing natural language processing (NLP) to utilize textual data more extensively. We analyze data from police stops in Philadelphia, using the full text of police reports to predict contraband discovery during "Terry stops."

We compare contraband predictions made with numerical and categorical data to those incorporating text data, focusing on the impact of race. Our findings show that including NLP and text data significantly alters the perceived biases in policing:

- Without text data: Significant bias against female suspects (-4.62 percentage points, $p = 0.048$), in favor of Black suspects (+1.31 percentage points, $p = 0.015$), and near-zero bias against Latino suspects.
- With text data: Near-zero bias for Black suspects (+0.02 percentage points, $p = 0.952$), reduced anti-female bias (-1.53 percentage points, $p = 0.386$), and increased anti-Latino bias (-1.41 percentage points, $p = 0.010$).

These results suggest that text matters significantly in assessing policing bias. Failing to control for text can produce misleading perceptions of bias or lack thereof. Our analysis implies that empirical studies of policing practices should incorporate methods that account for free-form text, potentially challenging earlier findings on bias that omit this crucial data.

Submitted to NeurIPS 2024 Workshop on Causality and Large Models (CaLM). Do not distribute.

2 Background

2.1 Setting

Our study analyzes data on pedestrian stops conducted by the Philadelphia Police Department between 2014 and 2023. The data were collected in an ongoing monitoring process stemming from the settlement agreement in the case of *NAACP v. City of Philadelphia*.

The Philadelphia Police Department provided data from police reports that occur after stops. Certain of the variables—like whether the police had reasonable suspicion for stops and frisks—were manually coded by lawyers as part of the monitoring process, on a randomly selected sample. We supplement the police data with demographic, economic, and crime data from the U.S. Census and the Philadelphia Police Department for additional controls.

The randomly selected sample comprises a total of 67,469 pedestrian stops (randomly selected from the full dataset), which served as the primary dataset we studied for this paper. Complete details about the specific variables we consider are below. We intend to conduct additional analysis on the full dataset, which has closer to a million observations, and when necessary will synthetically generate reasonable suspicion variables using fine-tuned large language models (as described below).

2.2 Prior Literature

Contraband discovery rates are an important outcome variable in empirical studies on policing, particularly in studying potential racial bias. Knowles et al. (2001) proposed an "outcome test," building on earlier work by Becker (1957), under which an unbiased police officer should find contraband on suspects of different races, genders, etc. at equal rates. Different rates of contraband discovery suggest bias insofar as "officers driven by racial prejudice will continue to search minority citizens at higher rates despite finding less contraband" (Tillyer and Klahm, 2011).

Studies conducted on this theoretical foundation have had mixed results. Some research has found no evidence of bias, with statistically equivalent contraband discovery rates between Black and White citizens (Knowles et al., 2001; Persico and Todd, 2006; Hernandez-Murillo and Knowles, 2004). However, other studies have found lower contraband discovery rates for minority citizens, suggesting potential bias in search practices (Engel and Johnson, 2006; Ridgeway, 2007). Tillyer and Klahm (2011) found that Black citizens were twice as likely as White citizens to be found with contraband in discretionary searches, suggesting reverse bias in favor of Black suspects (at least in discretionary searches; they found equal rates for mandatory searches).

Critics have raised concerns about the assumptions and limitations of using contraband discovery rates as evidence of bias. Most prominently, various critics have observed the specter of omitted variable bias—for a variety of reasons, the circumstances of police stops may differ depending on the race of the suspect (Anwar and Fang, 2006; Engel, 2008; Engel and Tillyer, 2008).

Police departments often have extensive records of the circumstances of police stops that could in theory be used to mitigate omitted variable bias. However, police records are natural language not easily converted to structured data. This article addresses this issue by using natural language processing, specifically large language models (LLMs), to incorporate natural language data in statistical analysis of contraband discovery.

3 Data and Methodology

3.1 Data

Our data come from the Philadelphia Police Department and describe pedestrian police stops and frisks from the years 2016 to 2023. Our sample of 67,469 total observations was randomly selected from over a million stops in this time period. A large number of variables are available, including information on outcomes, subjects, officers, and importantly, free text data. Table 1 includes summary statistics on a selection of the variables.

The data include information on whether a subject was arrested, an individual was frisked, whether contraband was discovered, and whether there was reasonable suspicion for the stop or the frisk. There is also information on the type of contraband recovered, including whether it was a gun or other weapon, drugs or something else. There is a great deal of data about the subject, including several variables about individual appearance, race, gender, age, and Latino status.

There is information about the location and time of the stop as well as identifiers for the officer and partner making the stop. We code location based on the Police Service Area (PSA) in which the stop was made. Crucially, there is a detailed free-text narrative by the police officer explaining the reason for the stop, which is intended to convey evidence of reasonable suspicion. The same information is also available for a frisk if one was made.

Table 1 contains summary statistics for each of the data we used in our analysis.

3.2 Methods

Applying related research that one of us is currently conducting, we use a new method to incorporate natural language in causal inference by leveraging predictions generated from a fine-tuned large language model (LLM). Specifcally, we fine-tune an LLM to generate direct predictions of the outcome variable, which are then used as an additional control in OLS regression.

Mathematically, a conventional OLS regression would take the form:

$$Y_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i + \varepsilon_i \tag{1}$$

Where Y_i is the dependent variable for observation i, β_0 is the intercept term, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients for the independent variables, $\mathbf{X}_i$ is a $k \times 1$ vector of independent variables for observation i, and ε_i is the error term for observation i.

We simply add an additional term to this regression:

$$Y_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{X}_i + \delta P_i^{LLM} + \varepsilon_i \tag{2}$$

Where δ is the coefficient for the LLM-predicted probability, and P_i^{LLM} is the predicted probability generated by the fine-tuned LLM for observation i.

We generate outcome predictions using Llama 3. Bai et al. (2023) show that transformers can learn various statistical models in context. Thie suggests that a fine-tuned LLM might theoretically be able to adapt textual inputs to a wide range of functional forms with sufficient training. We fine-tune Llama 3 using LoRA.[1]

To be specific, we trained Llama 3 to predict whether a suspect was discovered to have contraband using the stop narrative produced by the police officer. Then, once we fine-tuned Llama 3 to predict whether contraband was discovered, we incorporated the predicted probabilities generated by the LLM[2] as an additional variable in OLS regression. This allows us to analyze the residual variation in our predictive task while relying on OLS assumptions familiar to empirical legal scholars, simply adding an additional control.

We calculate the predictive performance of the OLS model, whether incorporating textual predictions or not, by using R-squared statistics as well as Mean Squared Error (MSE), testing MSE by training the OLS model on the training set and then calculating MSE on the test set.

Officer descriptions of stop events often include details that might be problematic for our analysis. Many of the descriptions describe demographic variables about the suspect that are already included in the OLS regression (for example, "Male was found..."), risking proxying and collinearity. In addition, many of the descriptions contain details not only of the events leading up to the stop but also the outcome of the stop (for example, "Suspect was released."). These muddy our predictive exercise, because it is very easy to predict that contraband is retrieved if the police report explicitly says so, which would remove important residual variation from our sample. If we had fine-tuned a model on the original dataset and found that coefficients on variables of interest had decreased, that could merely be a result of excessively detailed description.

For the fine-tunes we conducted that included natural language data, we pre-processed the datasets to allay the above concerns. For both datasets, we removed any explicit mention of demographic

[1]We use an 80%/20% training/testing split.

[2]These were straightforwardly generated by exponentiating the LLM's log probabilities

variables of interest from the police reports (for example, replacing "male suspect" with "individual"). Then we generated a version of the dataset redacting any mention of outcomes, and including only the information that would have been available to the office prior to making the stop (for example, simply removing the sentence "Suspect was released."). All edits to the reports were made using GPT-4o.

3.3 Hypotheses

Applying the methods described above, we test two different hypotheses in this paper.

First, we test the hypothesis that individual characteristics (for example, race and gender) will significantly influence the predicted values of dependent variables. If individual characteristics strongly predict outcomes, that may suggest police bias. As noted above, if Black suspects are less likely to be found with contraband than White suspects, that would suggest that the police are biased against Black suspects (because they are more likely to stop them even when unwarranted). Or, if Black suspects are more likely to be frisked after being stopped compared to White suspects, even holding the stated circumstances of the stop constant, that would again suggest that the police are biased against Black suspects. Thus regression analysis helps to shed light on potential policing biases.

Second, we test the hypothesis that text matters. By conducting prediction both with structured data only (excluding textual data) and with all data, including textual data, we can assess how important textual data is in the story and how well regressions conduct *ceteris paribus* analysis absent NLP. This point is important because virtually all analysis to date has occurred using categorical variables generated from textual data, rather than from textual data themselves; if conventional categorical variables are inadequate, that casts doubt on a huge swath of the literature and raises the inclusion of textual data as an important best practice for future work.

4 Analysis

4.1 Regression Equations and Model Performance

We conducted three OLS regressions using structured, non-textual data, and one regression including LLM predictions from textual data as an additional control. The regression equations are presented in Subsection A.3 of the Appendix. Table 2 and Figures 1 and 2 present performance statistics for each of the regressions. They show that R-squared and MSE dramatically improve when adding predictions based on natural language, suggesting that important residual variation is captured when these data are utilized.

4.2 Examples of Changed Predictions

The inclusion of NLP in our regression analysis led to significant changes in contraband prediction probabilities for individual cases. Here are two illustrative examples:

- **Example 1:**
  Police Report: "Radio call for a theft in progress ... Individual matched [description and] was observed carrying [stolen items]. Loss prevention officer ... positively identified the suspect as the person who stole the [items]."

LPM probability without NLP:	-2.494%
LPM probability with NLP:	**44.902%**

  In this case, the NLP model significantly increased the predicted probability of contraband discovery. The detailed description of a theft in progress, along with positive identification by a loss prevention officer, likely contributed to this substantial increase.

- **Example 2:**
  Police Report: "The suspect was observed by police driving ... and failed to use a right turn signal..."

LPM probability without NLP:	39.258%
LPM probability with NLP:	**3.186%**

Conversely, in this example, the NLP model dramatically decreased the predicted probability of contraband discovery. The report describes a minor traffic violation, which provides an explanation for the stop that makes contraband discovery unlikely (although still not impossible), compared to the pre-NLP model.

These examples demonstrate how the inclusion of textual data through NLP can lead to more nuanced and context-aware predictions, correcting for biases or oversimplifications in models relying solely on structured data.

4.3 Main Results

Table 3 and Figures 3, 4, 5, and 6 present the results of the regression analysis. The model including only structured data suggests that female suspects were 4.62 percentage points less likely to be found with contraband than male suspects ($p = 0.048$); Black suspects were 1.31 percentage points more likely to be found with contraband than White suspects ($p = 0.015$); Asian suspects were 1.20 percentage points more likely to be found with contraband than White suspects ($p = 0.598$); and Latino suspects were 0.37 percentage points less likely to be found with contraband than non-Latino suspects ($p = 0.607$).

Under the conventional literature on contraband discovery, these results suggest strong anti-female bias, moderate pro-Black and pro-Asian bias, and near-zero Latino bias. Only the results regarding anti-female and pro-Black bias were significant at the 95% level.

However, when textual data are included in the training and test datasets, coefficient estimates dramatically shift in magnitude. When controlling for textual data, female suspects were only 1.53 percentage points less likely to be found with contraband than male suspects ($p = 0.386$); Black suspects were only 0.02 percentage points more likely to be found with contraband than White suspects ($p = 0.952$); Asian suspects were 1.43 percentage points more likely to be found with contraband than White suspects ($p = 0.403$); and Latino suspects were 1.41 percentage points less likely to be found with contraband than non-Latino suspects ($p = 0.010$).

This suggests near-zero bias regarding Black suspects, substantially less (and statistically insignificant) bias regarding females versus males, and definite bias against Latino suspects, which is also statistically significant at the 99% level.[3]

In summary, while the models trained only on structured data show a variety of police biases on the demographic variables we tested, the inclusion of free-form textual data dramatically changes those estimates of bias, moving us from an estimate of anti-female and pro-Black bias to an estimate of anti-Latino bias. This suggests that the biases observed with structured data alone may be mitigated or complicated by additional contextual information captured in text.

4.4 Limitations and Robustness Checks

4.4.1 Biased Police Report Descriptions

Because so much of our analysis relies on descriptions written by police officers who we hypothesize may have bias, a natural concern is that the descriptions themselves exhibit bias. It could be, for example, that when police see a suspect smoking something hand-rolled, they might describe it as "likely tobacco" if the suspect is White and "likely marijuana" if the suspect is Black. Or, to take another example, police might be concerned about being accused of gender bias and therefore devote extra care to making their report sound suspicious when stopping a woman. In either case, bias in natural language descriptions would serve as a confounder in our analysis.

We can test this possibility in the OLS regression that incorporates predicted probabilities, by interacting the predicted probability with each of the variables of interest—i.e., generating an interaction between the indicator variable for "Female", "Black", etc. with the predicted probabilities. Table 3 shows the results of this analysis. The coefficients on the interactions for female and Black suspects are statistically insignificant ($p = 0.317$ and $p = 0.514$, respectively), the coefficient for Latino suspects is significant at 90% but not 95% ($p = 0.056$), and each of the aforementioned coefficients is

[3]The bias estimates regarding Asian suspects are similar in magnitude but noisier–in addition, there is the possibility of bias in these estimates, discussed below.

near zero in magnitude (-0.1244, -0.0133, and -0.0493, respectively; note that this term is a multiplier against the value of the prediction coefficient which itself has a mean value of 0.075). However, the coefficient for Asian suspects is very significant (p = 0.000) and large in magnitude (0.4088).

The positive sign on the coefficient for Asian suspects implies that predictions based on police reports alone underestimate the likelihood that Asian suspects will be found to have contraband. Incorporating an NLP control therefore introduces upward bias on the coefficient for Asian suspects–that is, relative to other suspects, the model interprets Asian suspects as having contraband at higher rates by virtue of being Asian rather than by virtue of the omitted variable that causes bias in the police reports. This means that the police may be more biased against Asian suspects than the NLP controls suggest.

The fact that this interaction term is significant does not, however, tell us anything about the *cause* of the reporting bias. One possibility is that the police might be reporting facts differently depending on the race of the suspect, making Asians seem less suspicious than suspects of other races *ex post facto*. Alternatively, it might be that police are accurately and evenhandedly reporting facts, but that on the same set of facts Asian suspects are more likely to be carrying contraband.

4.4.2 Just-So Reports and Coefficient Attenuation

Even if police reports are unbiased, controlling for their content could inappropriately attenuate coefficient estimates in contraband discovery if the police tend to write just-so reports, reshaping the narrative in their police reports after the fact to make the stop seem justifiable.

Note that this is not a concern if police merely engage in across-the-board puffery–for example, if the police were always to make events sound 50% more convincing than reality, the fine-tuned LLM would account for this, since its predictions are ultimately rooted in actual contraband results and it would simply discount for the 50% puffery in its predicted probabilities. Because our dataset consists only of cases where stops were made, police have a consistent incentive to give the appearance of reasonable suspicion, which would tend to give rise to level bias controllable by the LLM's training.

But the attenuation problem remains when different sorts of stops are differently misreported. Here, too, certain directions of misreporting are less problematic. If police were to take greater care to make stops seem justified when no contraband is ultimately found (because they might think the contraband speaks for itself in cases where it is found), this would simply reduce the predictive power of the model and make it a less effective control. On the other hand, if police were to distort their reporting to make stops seem more justified in cases where contraband was found (i.e. in cases where stops really *were* justified), that would essentially turn the LLM control into an "over-control" and lead to attenuation of the magnitude of other coefficients in the OLS regression.

This possibility is more fundamental and more difficult to test. There is some evidence that police either do not try to or are not very good at mis-reporting in general, like the relative frequency with which police make stops that are later judged to lack reasonable suspicion (20.5%), and much of the existing literature (which often extracts simple controls from stop narratives) operates under the same assumption that stop narratives accurately reflect what happened. Moreover, the fact that we do not see attenuation across the board (the coefficient for the Latino indicator variable increases in magnitude) is some evidence against the potential for attenuation from just-so reporting. However, the possibility remains.

5 Conclusion

This paper demonstrates the significant impact that NLP techniques can have when analyzing police stop data for potential racial bias. By leveraging the full text descriptions provided by officers rather than just numerical and categorical data, NLP methods can produce substantially different results, in this case causing apparent police biases to disappear. This finding highlights the importance of considering free-form text in analyses of policing practices and casts doubt on some prior conclusions regarding bias that did not incorporate such textual information.

References

Shamena Anwar and Hanming Fang. 2006. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review*, 96(1):127–151.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. In *37th Conference on Neural Information Processing Systems*.

Gary S Becker. 1957. *The economics of discrimination*. University of Chicago Press.

Robin S Engel. 2008. Revisiting the use of citizen demeanor by police officers: An empirical assessment. *Crime Delinquency*, 54(4):600–636.

Robin S Engel and Rob Tillyer. 2008. A critique of the "outcome test" in racial profiling research. *Justice Quarterly*, 25(1):1–36.

Robin Shepard Engel and Richard Johnson. 2006. Toward a better understanding of racial and ethnic disparities in search and seizure rates. *Journal of Criminal Justice*, 34(6):605–617.

Ruben Hernandez-Murillo and John Knowles. 2004. Racial profiling or racist policing? bounds tests in aggregate data. *International Economic Review*, 45(3):959–989.

John Knowles, Nicola Persico, and Petra Todd. 2001. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229.

Nicola Persico and Petra E Todd. 2006. Racial profiling? detecting bias using statistical evidence. *Annual Review of Economics*, 1:229–254.

Greg Ridgeway. 2007. Analysis of racial disparities in the new york police department's stop, question, and frisk practices. *RAND Corporation*.

Rob Tillyer and Charles F Klahm. 2011. Searching for contraband: Assessing the use of discretion by police officers. *Police Quarterly*, 14(2):166–185.

A Appendix

A.1 LLM Fine-Tuning Hyperparameters

We used the following hyperparameters and configurations to fine-tune Llama 3:

- **Base Model:** We used a 4-bit quantized version of the Llama 3 8B Instruct model, which allows for faster loading and reduced memory usage.
- **Sequence Length:** The maximum sequence length was set to 2048 tokens.
- **Quantization:** We employed 4-bit quantization to reduce memory usage and enable faster training.
- **LoRA Configuration:**
  - Rank (r): 16
  - LoRA Alpha: 32
  - LoRA Dropout: 0
  - Bias: "none"
- **Training Configuration:**
  - Batch Size: 16 per device
  - Gradient Accumulation Steps: 1
  - Warmup Steps: 100
  - Number of Epochs: 3
  - Learning Rate: 0.0001
  - Optimizer: AdamW (8-bit)

    - Weight Decay: 0.01
    - Learning Rate Scheduler: Cosine
- **Precision:** We used mixed precision training, automatically selecting between FP16 and BF16 based on hardware support.
- **Gradient Checkpointing:** We used Unsloth for gradient checkpointing.

A.2 Prompt to Remove Outcome Language from Police Reports

We used gpt-4o-2024-05-13, with 4096 max tokens and temperature 0.000001. We used a single prompt for all redactions:

```
You are being given the contents of a police report describing
the events of a police stop.  You have the following jobs:  1.
Remove the following demographic information:  (a) the race
of the suspect (e.g.  convert "Black individual found..." to
"Individual found...") (b) the gender of the suspect (e.g.
convert "Male was found..." to "Suspect was found...") (c)
whether the suspect was Hispanic/Latino or not (e.g.  convert
"Hispanic individual found..." to "Individual found...") (d)
the age of the suspect (e.g.  convert "Young suspect found..."
to "Suspect found...") 2.  If the report is in all-caps,
convert it into sentence case.  3.  Remove any discussion
of the outcome of the police stop.  Leave only the information
the police would have known before making the stop, and delete
any information about what transpired after the police stopped
the suspect.  4.  Return only the modified text, without any
additional explanations or comments.  If no text is given,
just reply "N/A".
```

A.3 Regression Equations

$$\text{Contraband}_i = \beta_0 + \beta_1 \text{Female}_i + \sum_r \beta_r \text{Race}_{ri} + \beta_2 \text{Latino}_i + \varepsilon_i \tag{3}$$

$$\begin{aligned}\text{Contraband}_i = &\beta_0 + \beta_1 \text{Female}_i + \sum_r \beta_r \text{Race}_{ri} + \beta_2 \text{Latino}_i + \beta_3 \text{Age}_i + \beta_4 \text{Height}_i + \beta_5 \text{Weight}_i \\ &+ \sum_b \beta_b \text{Build}_{bi} + \sum_f \beta_f \text{FH}_{fi} + \beta_6 \text{WithPartner}_i + \beta_7 \text{MinSince2000}_i \\ &+ \sum_m \beta_m \text{Month}_{mi} + \sum_t \beta_t \text{ToD}_{ti} + \varepsilon_i \end{aligned} \tag{4}$$

$$\begin{aligned}\text{Contraband}_i = &\beta_0 + \beta_1 \text{Female}_i + \sum_r \beta_r \text{Race}_{ri} + \beta_2 \text{Latino}_i + \beta_3 \text{Age}_i + \beta_4 \text{Height}_i + \beta_5 \text{Weight}_i \\ &+ \sum_b \beta_b \text{Build}_{bi} + \sum_f \beta_f \text{FH}_{fi} + \beta_6 \text{WithPartner}_i + \beta_7 \text{MinSince2000}_i \\ &+ \sum_m \beta_m \text{Month}_{mi} + \sum_t \beta_t \text{ToD}_{ti} + \sum_p \beta_p \text{Officer}_{pi} + \sum_s \beta_s \text{PSA}_{si} + \varepsilon_i \end{aligned} \tag{5}$$

$$\begin{aligned}\text{Contraband}_i = &\beta_0 + \beta_1 \text{Female}_i + \sum_r \beta_r \text{Race}_{ri} + \beta_2 \text{Latino}_i + \beta_3 \text{Age}_i + \beta_4 \text{Height}_i + \beta_5 \text{Weight}_i \\ &+ \sum_b \beta_b \text{Build}_{bi} + \sum_f \beta_f \text{FH}_{fi} + \beta_6 \text{WithPartner}_i + \beta_7 \text{MinSince2000}_i \\ &+ \sum_m \beta_m \text{Month}_{mi} + \sum_t \beta_t \text{ToD}_{ti} + \sum_p \beta_p \text{Officer}_{pi} + \sum_s \beta_s \text{PSA}_{si} \\ &+ \beta_8 \text{PredictedContraband}_i + \varepsilon_i \end{aligned} \tag{6}$$

A.4 Tables and Figures

Table 1: Summary statistics for variables in the dataset.

Variable	Mean	SD	Variable	Mean	SD
CONTRABAND_DISCOVERY_No	0.873	0.333	RACE_Unknown	0.007	0.084
CONTRABAND_DISCOVERY_Yes	0.069	0.254	LATINO_No	0.904	0.294
STOPRS_No	0.147	0.354	LATINO_Yes	0.096	0.294
STOPRS_N/A	0.143	0.350	AGE	33.268	13.274
STOPRS_Yes	0.570	0.495	HEIGHT	5.517	0.425
GENDER_Male	0.859	0.348	WEIGHT	170.543	33.598
GENDER_Female	0.141	0.348	BUILD_Thin	0.354	0.478
RACE_Black	0.702	0.458	BUILD_Heavy	0.069	0.253
RACE_White	0.281	0.450	BUILD_Medium	0.445	0.497
RACE_Asian	0.010	0.097	BUILD_Tall	0.015	0.123
BUILD_Small	0.025	0.157	FACIAL_HAIR_Unshaven	0.244	0.430
BUILD_Thin,Small	0.006	0.077	FACIAL_HAIR_Beard	0.282	0.450
BUILD_Stocky	0.034	0.182	FACIAL_HAIR_Goatee	0.089	0.284
BUILD_Medium,Thin	0.005	0.074	FACIAL_HAIR_Mustache	0.056	0.231
BUILD_Thin,Tall	0.013	0.113	minutes_since_2000	9935994.119	1262457.312
BUILD_Muscular	0.006	0.077	month	5.891	3.052
time_of_day_Evening	0.566	0.496	WITH_PARTNER_No	0.256	0.436
time_of_day_Day	0.314	0.464	WITH_PARTNER_Yes	0.744	0.436
time_of_day_Night	0.120	0.325	predicted_contraband	0.075	0.176

Table 2: Comparison of R-squared and Mean Squared Errors (MSE) for different regression types

Regression Type	R-squared		MSE	
	Train	Test	Train	Test
Key Variables	0.0026	0.0023	0.0679	0.0671
Key + Basic Control	0.0188	0.0198	0.0662	0.0651
All Structured Variables	0.1586	0.0584	0.0568	0.0625
All Structured Variables + Predicted	0.5062	0.3368	0.0333	0.0440

Table 3: Regression Results for Contraband Discovery

	Key Variables	**Key + Basic Control**	**All Struct. Vars**	**All Struct. + Predicted**	**All Struct. + Interact.**
Female	-0.0300***	-0.0580**	-0.0462**	-0.0153	-0.0098
	(0.0041)	(0.0233)	(0.0233)	(0.0176)	(0.0184)
Black	0.0141***	0.0090*	0.0131**	0.0002	0.0014
	(0.0034)	(0.0047)	(0.0054)	(0.0041)	(0.0044)
Asian	0.0032	0.0077	0.0120	0.0143	-0.0125
	(0.0148)	(0.0225)	(0.0227)	(0.0171)	(0.0184)
Latino	0.0237***	0.0169**	-0.0037	-0.0141***	-0.0092
	(0.0053)	(0.0068)	(0.0072)	(0.0055)	(0.0060)
Female × Pred.					-0.1244
					(0.124)
Black × Pred.					-0.0133
					(0.020)
Asian × Pred.					0.4088***
					(0.103)
Latino × Pred.					-0.0493*
					(0.026)

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

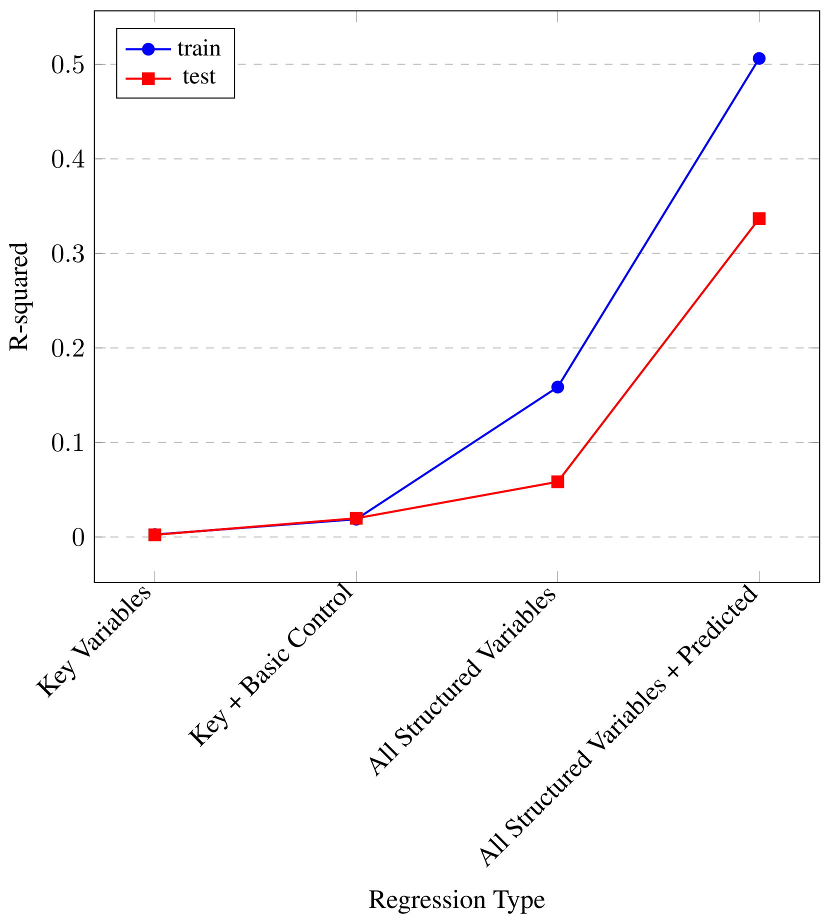

Figure 1: Comparison of R-squared for training and test sets depending on variables included in OLS regression.

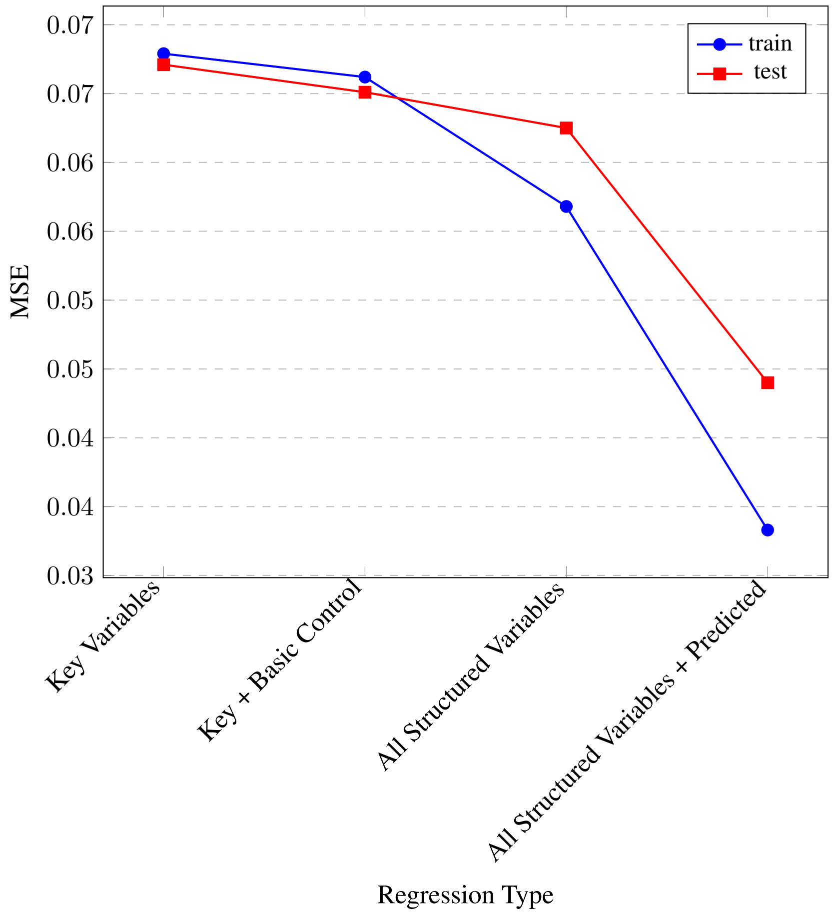

Figure 2: Comparison of Mean Squared Errors for training and test sets depending on variables included in OLS regression.

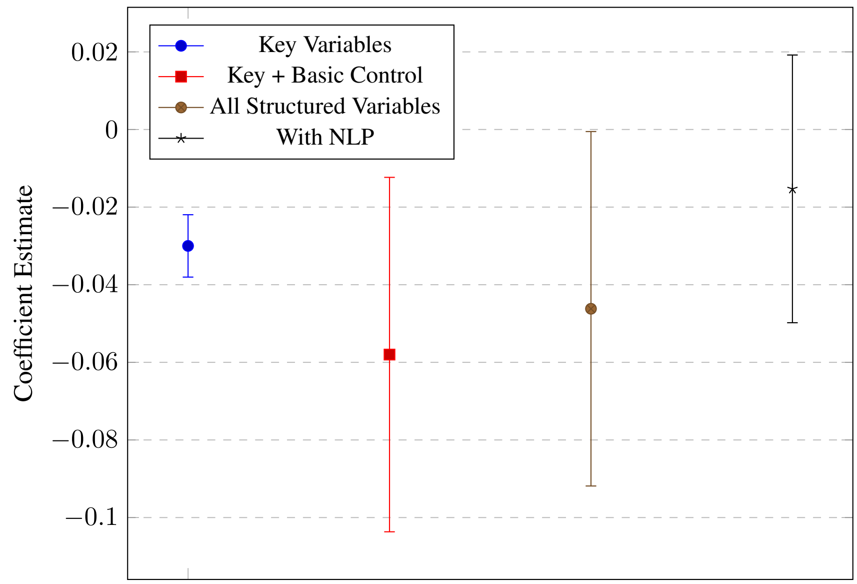

Figure 3: Coefficient estimates for female indicator variable from different regression models, with 95% confidence intervals

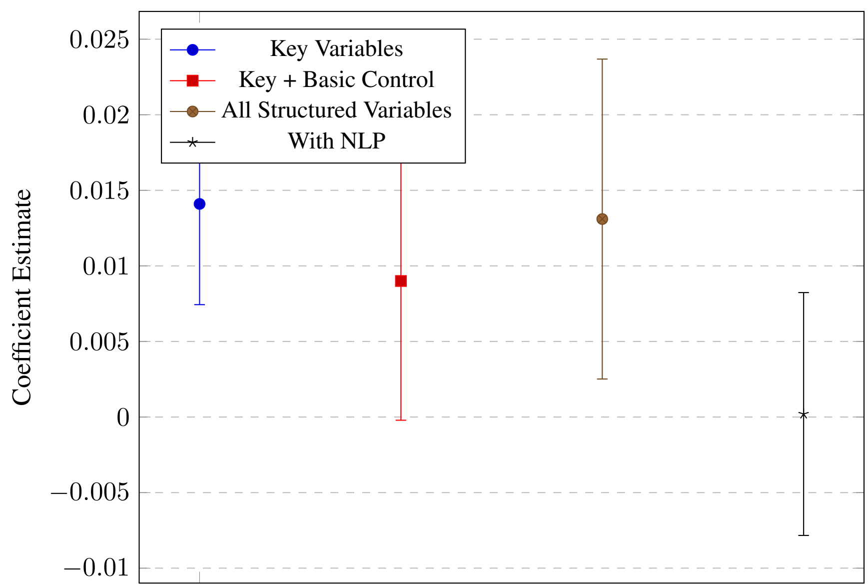

Figure 4: Coefficient estimates for Black indicator variable from different regression models, with 95% confidence intervals

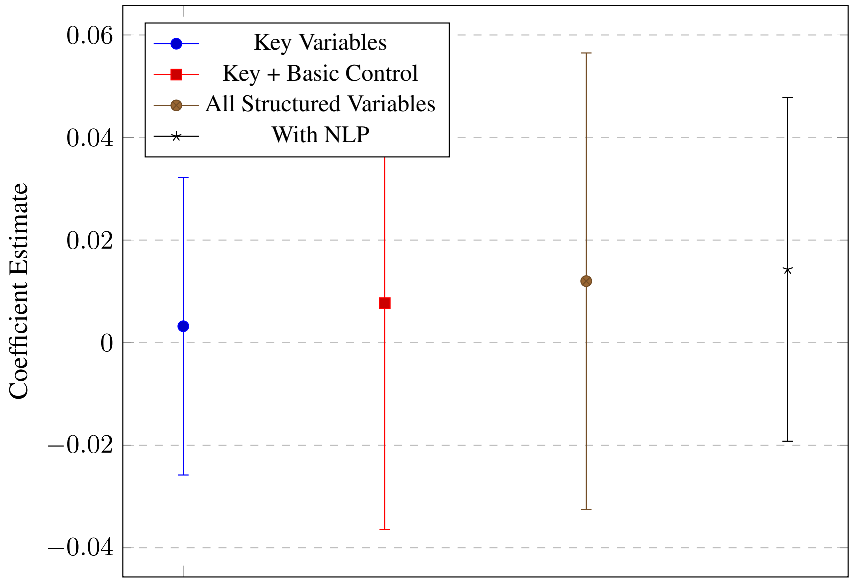

Figure 5: Coefficient estimates for Asian indicator variable from different regression models, with 95% confidence intervals

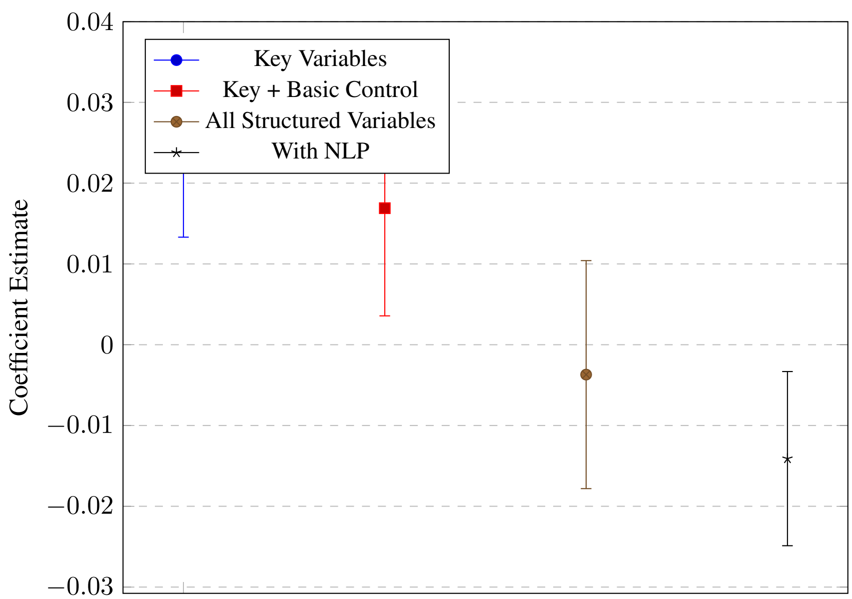

Figure 6: Coefficient estimates for Latino indicator variable from different regression models, with 95% confidence intervals