Seattle Office of Inspector General

Evaluation of Disparity Review Methodology

An Independent Evaluation of Statistical Methods Utilized by the Seattle Police Department

Prepared by
Miroslava Meza, M.S.
John Campos, M.A.

March 18, 2019

Office of Inspector General
City of Seattle
PO Box 94764
Seattle, WA 98124-7064
oig@seattle.gov
(206) 684-3663

Seattle Office of
Inspector General

Summary

The Office of Inspector General for Public Safety (OIG) conducted an evaluation of the methodology proposed by the Seattle Police Department (SPD) to analyze potential racial disparity in investigative stops,[1] contained in the *SPD Disparity Review – Part 1* report.[2]

SPD planned to use a rigorous statistical method called Propensity Score Matching (PSM) to determine whether certain race/ethnicity groups were disproportionately represented in investigative stops when compared to characteristics of the subject of the stop, officer, or the event.

PSM is less biased than other common statistical approaches used to measure disparity, because it contains inherent procedures that reduce the proliferation of data manipulation errors and the personal choices of the analyst when selecting the paired-samples for modeling.

The OIG evaluation consisted of (1) a literature review to compare four statistical methods and their potential applicability to the SPD dataset;[3] and (2) empirical testing of the four statistical methods using SPD disparity data.

OIG concludes that PSM is an appropriate statistical method for identifying disparity with regard to the race/ethnicity of the subject stopped when compared to various characteristics of the subject, officer or event, given the type and quality of SPD data. PSM contains inherent procedures that reduce the proliferation of data manipulation errors and the personal choices of the analyst when selecting paired-samples for modeling.

[1] Also known as Terry Stops.

[2] The Consent Decree (Dkt. 3-1 ¶ 145), *United States of America v. City of Seattle*, 12 Civ. 1282 (JLR), states that SPD "should deliver police services that are equitable, respectful, and free of unlawful bias, in a manner that promotes broad community engagement and confidence in the Department." SPD Policy 5.140(9) for disparate impacts[2] notes that, in consultation with OIG, SPD shall periodically analyze data, including stops, that may have a disparate impact on particular protected classes.

[3] Propensity Score Matching (PSM), least squares regression, logistic regression and logistic regression with blocked paired-sampling.

Issue

SPD Disparity Review

The *SPD Disparity Review – Part 1* report estimates whether there are racial/ethnic disparities in policing practices affecting protected classes.

The information available to evaluate the disparity between white and non-white subjects stopped for investigation is a mix of data SPD collects[4] regarding the events, officers, and the subjects of the stops (see Table 1). See the Frequencies Table in the Appendix for a list and frequencies of the variables used to describe these characteristics.

Table 1. Variables, Types and Classifications as registered in SPD dataset.

Variables	Type	Classification
Dependent[5]		
Subject's Race/Ethnicity group[6]	Dichotomous	Subject
Independent[7]		
Date	Interval	Event
Time	Interval	Event
Initial Call Type	Ordinal	Event
Priority	Ordinal	Event
Subject's Age at Contact	Interval	Subject
Subject's Gender	Categorical	Subject
Officer's Age at Contact	Continuous	Officer
Officer's Gender	Categorical	Officer
Officer's Race	Categorical	Officer
Officer's Title	Ordinal	Officer
Officer's Years of Service	Ordinal	Officer
Officer's CIT-Certified	Dichotomous	Officer
Officer's Assignment (Squad)	Ordinal	Officer

[4] SPD officers in the field enter into their patrol car's terminal information for each investigative stop during their shift. If a stop does not result into a detention, the subject of the stop is not obliged to provide any personal data. It is up to the officer performing the stop to determine the characteristics of the subject stopped. That information is stored in SPD databases.

[5] Dependent Variable is the outcome we are measuring, which is the race/ethnicity of a subject stopped for an investigative stop.

[6] 1=(American Indian – Alaska Native, Asian, Black, Hispanic, Others, Non-White) and 0=(White).

[7] Key pieces of data from the investigative stops that may influence the race/ethnicity of people subject to investigative stops.

SPD seeks to answer two questions about disparity in investigative stops:

1. Is there disparity between the following race/ethnicity groups? *Non-White vs. White, American Indian – Alaska Native vs. White, Asian vs. White, Black vs. White, Hispanic vs. White, Others*[8] *vs. White*
2. Within these race/ethnicity groups, what is the disparity in investigative stops, if any, by certain characteristics of the event, officer or subject?

There are challenges in answering these questions such as:

1. There are a limited number of statistical methods that model the response of a dichotomous variable due to a mix of independent variables (predictors) of different types – continuous, discrete, categorical, interval, ordinal. [9]
2. Certain characteristics of the subjects are assumed by the officer. For example, the subject's gender, age and race/ethnicity are the officer's assumption.
3. The analyst may introduce selection bias by selecting a sample that does not properly estimate racial/ethnic disparity in police practices.

Taking these challenges into consideration and desirable characteristics for a reliable disparity analysis method, SPD proposes to use PSM.

OIG Analysis

OIG conducted an independent review of proposed SPD methods for analyzing disparity in investigative stops to answer the following questions:

1. Is PSM an appropriate statistical test, given the characteristics of the data?
2. Is there a valid alternative test that could be used?

[8] The race/ethnicity group labeled as "other" includes those subjects the officer conducting the investigative stop could not identify as part of one of the existing categories.

[9] The characteristics used to identify potential predictors of disparity are not measured in the same units. For example, officer years of service are recorded in integers, the types of calls that originated the stops are recorded using a scale from one to four, while the age of the investigative stop subjects are estimated by the officers and recorded as an age range.

Findings

Literature Review

In order to effectively evaluate SPD statistical methods for analyzing investigative stops, OIG reviewed current literature concerning the general characteristics and applicability of PSM[10-13] and other statistical techniques, in order to compare methods and potential applicability.[14-21]

Upon review, OIG ascertained that methods using logistic *regression*[22] are adequate to identify disparity with regard to the race/ethnicity of the subject stopped. Logistic regression estimates the effects on dichotomous[23] dependent variables (e.g., race/ethnicity).

[10] Heckman JJ, Todd PE. A note on adapting propensity score matching and selection models to choice based samples. National Bureau of Economic Research (NBER). Working Paper Series, Working Paper 15179, 2009.

[11] Ho D, Imai K, King G, Stuart E. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis 15(3): 199-236, 2007.

[12] Ho D, Imai K, King G, Stuart E. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, 2007. http://gking.harvard.edu/matchit/ (Accessed February 20, 2019).

[13] Zhang Z, Kim HJ, Lonjon G, Zhu Y; written on behalf of AME Big-Data Clinical Trial Collaborative Group. Balance diagnostics after propensity score matching. Ann Transl Med 7(1):16, 2019.

[14] Hoffmann, John P. Regression Models for Categorical, Count and Related Variables: An Applied Approach. University of California Press (2016). URL: https://www.jstor.org/stable/10.1525/j.ctv1wxrfr.6 (Accessed February 27, 2019).

[15] Kirk RE. Experimental design: Procedures for the behavioral sciences. Thousand Oaks, CA: SAGE Publications, Inc., 2013.

[16] MacDonald, JM. Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. Journal of the American Statistical Association. 104:486, 661-668, 2009.

[17] Morris S, Dunleavy E (eds). Adverse impact analysis: Understanding data statistics and risk. New York: Routledge, 2016.

[18] Ridgeway G. Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. Journal of quantitative criminology, 22(1): 1-29, 2006.

[19] Rosenbaum PR. Observation and experiment: an introduction to causal inference. Cambridge, Massachusetts: Harvard University Press, 2017.

[20] Williams MN, Grajales CAG, Kurkiewicz D. Assumptions of multiple regression: Correcting two misconceptions. Practical Assessment, Research & Evaluation, 18(11), 2013.

[21] Zuberi T, Bonilla-Silva E. (eds.) White Logic, White Methods: Racism and Methodology. Lanham: Rowman & Littlefield Publishers, 2008.

[22]PSM, as a statistical method, performs a series of steps to automatically select a sample that reduces for selection bias and it is followed by a logistic regression to estimate the interaction between dependent and independent variables.

[23] Dichotomous is also known as binary.

Additionally, the literature recommends using blocked paired-sampling when comparing two populations in order to approximate an experimental design (quasi-experimental). This approach helps reduce selection bias. That is, blocked paired-sampling reduces the possibility of selecting samples to measure disparity that are not representative of the population that is currently subject to investigative stops.

Table 2 compares the general characteristics of PSM, logistic regression and logistic regression with blocked paired-sampling. Ordinary Least Squares (OLS) does not appear in this table because the results of both literature review and empirical tests showed that OLS is insufficient to measure causal effects of disparity, as described in the empirical testing section below.

Table 2: General Characteristics from Literature Review	**Propensity Score Matching (PSM)**	**Logistic Regression**	**Logistic Regression with blocked paired-sampling**
Compares differences between two groups	Y	Y	Y
Adequate for categorical dependent variables (0/1)	Y	Y	Y
Adequate for a mix of independent and control variables of the following types: categorical, dichotomous, continuous, ranked, and factored	Y	Y	Y
Logistic Regression	Y	Y	Y
Quasi-experimental design	Y	**N**	Y
Appropriate for two populations where one is significantly smaller than the other	Y	**N**	Y

As can be seen in Table 2, of the methods explored, only PSM and logistic regression with blocked paired-sampling employ a quasi-experimental design. They also have the positive attribute that they can be used for comparing two populations even when one of them is significantly smaller than the other, which is the case with the SPD race/ethnicity variable.

Empirical Testing

Following the literature review and subsequent discussion with the key stakeholders in SPD and representatives of the federal monitoring team, OIG empirically tested three logistic regression methods and ordinary least squares regression, using a sample dataset provided by SPD.

1) Ordinary Least Squares Regression

Ordinary Least Squares (OLS) is a common estimation method when the following characteristics are assumed: linearity, independence of the prediction errors, independent sample, homoscedasticity, lack of autocorrelation among the errors, collinearity, and normal distribution of the errors.

The empirical test for OLS consisted in creating an OLS regression model with SPD data (see Table 3), followed by testing for the assumptions mentioned above. Since none of those assumptions are met using the SPD dataset, OLS is deemed inadequate.

Table 3. OLS			**Empirical Test Results**	
	OIG Evaluation Criteria	**General Case**	**SPD Data**	**Notes**
1	Appropriate for observational studies.	Some cases	N	Possible only if the analyst creates random samples of paired data or designs a quasi-experiment.
2	Capable of measuring disparity between two groups.	Y	Y	Uses paired data to calculate the Mean Causal Effect of the differences between the two groups (see Eq. 3 in Appendix).
3	Capable of handling a dichotomous dependent variable.	Some cases	N	Coefficient p-values >.05.
4	Capable of handling multiple independent variables of different types (dichotomous, ordinal, categorical, interval, count).	Some cases	N	Coefficient p-values>.05, and r^2<.001.
5	Preserves its power even when there is a significantly smaller sample of one of the observed groups.	N	N	Coefficient p-values>.05, and r^2<.001.
6	Reduces the effect of control variables that are not under analysis.	N	N	This method does not account for control variables.
7	Reduces the number of opportunities to introduce errors and bias due to data manipulation.	N	N	This method does not have any strategy to deal with errors and analyst selection bias.

2) Logistic Regression

Logistic regression allows for evaluation of effects between a dichotomous dependent variable and multiple independent variables of mixed types (e.g., continuous, discrete, categorical, ordinal, interval, and dichotomous). This model calculates the odds of being stopped for investigation when two race/ethnicity groups are compared.[24]

This method allows for comparison of the Mean Difference Effects (see Appendix for the disparity analysis explained as a random variable) between two treatments that are not randomly assigned (e.g., protected group vs. non-protected group). This approach may create a robust regression model when the protected group is compared to <u>all</u> the race/ethnicity-based non-protected groups in aggregate. This method requires labor-intensive manual matching. Without matching, it would generally analyze the entire dataset of investigative stops without blocking effects of variables that are not of interest.

[24] This technique compares two groups at a time, or two specific outcomes (e.g. Hispanic vs. White).

Limitations of logistic regression (when the sample used for the model is not treated for selection bias) are exposed when comparing non-protected subjects to small racial or ethnic protected groups. In this case, there are limitations in the ability of the sample to represent each race/ethnicity group and in the ability of the test to capture effects in multiple possible causes of disparity. As a result, most factors will appear to be statistically insignificant.

In this case, logistic regression analysis shows that groups differ, but the cause of the disparity will be ambiguous,[25] because most of the independent variables have a low count of occurrence and without an experimental or quasi-experimental design[26] the model does not identify their causal relationships (if any actually exist).

Table 4. Logistic Regression		**General Case**	**Empirical Test Results**	
OIG Evaluation Criteria			**SPD Data**	**Notes**
1	Appropriate for observational studies.	Some cases	N	Only if it is used as part of a quasi-experimental design.
2	Capable of measuring disparity between two groups.	Some cases	Y	Uses paired data to calculate the Mean Causal Effect of the differences between the two groups (see Eq. 3 in Appendix).
3	Capable of handling a dichotomous dependent variable.	Some cases	Y	Meets the assumptions needed.
4	Capable of handling multiple independent variables of different types (dichotomous, ordinal, categorical, interval, count).	Some cases	Y	Meets the assumptions needed.
5	Preserves its power even when there is a significantly smaller sample of one of the observed groups.	N	N	When comparing the largest group vs. the smaller (white, others) this model cannot measure the effects on most independent variables.
6	Reduces the effect of control variables that are not under analysis.	Some cases	N	This method does not account for control variables.
7	Reduces the number of opportunities to introduce errors and bias from data manipulation.	N	N	This method does not have any strategy to deal with errors and analyst bias.

3) Logistic Regression with Blocked Paired-Sampling

This variation of logistic regression reduces the effect of factors that are not under analysis and guarantees that smaller race/ethnicity groups have a match-sample for each subject. Logistic regression is performed on a blocked paired-sample (i.e., match) derived from all of the investigative stops. The analyst matches paired samples of subjects with very similar characteristics in their control variables (e.g., age, location, call type, etc.).

[25] Morris, Dunleavy, 2016.

[26] See Quasi-experimental approach section in this document for detailed explanation.

The analyst performs the matches using one of several techniques (e.g., nearest neighbor, full matching and optimal matching). By performing the logistic regression, the analyst can verify whether the blocked paired-sample is useful and if the assumptions of the model are met. If the assumptions are not met, it may be because the paired-sample chosen is not well-matched or because the independent variables are not associated to the race/ethnicity of the subject stopped. The analyst may then match the data again using other techniques until the analyst has a set that satisfies the logistic regression assumptions, or the analyst can look for another model to evaluate the effects of the variables.

Table 5. Logistic Regression with Blocked Paired Sample OIG Evaluation Criteria		**General Case**	**Empirical Test Results**	
			SPD Data	**Notes**
1	Appropriate for observational studies.	Y	Y	This is a quasi-experimental approach, appropriate for two treatments not randomly assigned.
2	Capable of measuring disparity between two groups.	Some cases	Y	Uses paired data to calculate the Mean Causal Effect of the differences between the two groups (see Eq. 3 in Appendix).
3	Capable of handling a dichotomous dependent variable.	Some cases	Y	Appropriate for dichotomous or binary dependent variables.
4	Capable of handling multiple independent variables of different types (dichotomous, ordinal, categorical, interval, count).	Some cases	Y	It can include different types of variables, if the assumptions are met.
5	Preserves its power even when there is a significantly smaller sample of one of the observed groups.	Y	Y	Using paired-datasets reduces issues of lack of representation of a group.
6	Reduces the effect of control variables that are not under analysis.	Y	Y	It uses a matched paired-sample that reduces the effect of control variables.
7	Reduces the number of opportunities to introduce errors and bias from data manipulation.	N	N	Sample matching is performed by an algorithm with the analyst fitting the selection criteria and data processing (matching) by trial and error.

4) Propensity Score Matching (PSM)

PSM is a statistical method with a logistic regression that estimates differences between two groups when it is not possible to randomly assign subjects to each group and afterwards observe the results. [27] PSM achieves this measurement by using all the captured information of investigative stops and matching the data in comparison groups. The PSM matching process automatically creates paired samples where the two matched sets have

[27] For example, in the SPD dataset, age, ethnicity, location, etc. are not randomly assigned by an experiment, but rather collected in the field.

very similar characteristics (e.g., age of the subject, age of the officer, location, time of the day, call type, etc.). The match is created by a programmatic algorithm.

In PSM, the analyst cannot assess or adjust the sample by looking to the output of the regression. This reduces the effect of external factors that are not under analysis, ensures that smaller race/ethnicity groups have a match-sample for each of their groups' subjects, and limits analyst selection bias.

PSM separates the estimation procedure into two steps. The first step simulates the research design of an experiment, where no information on outcomes is known. The second step consists of reviewing the fitness of the regression using mean covariances as proxies for fitness without looking at the causal model's results. Analyzing the results is a third, independent step where PSM uses this matched sample to perform a logistic regression that accounts for the effect of the independent variables on disparity.

Table 6. Propensity Score Matching		**General Case**	**Empirical Test Results**	
OIG Evaluation Criteria			**SPD Data**	**Notes**
1	Appropriate for observational studies.	Y	Y	PSM is a quasi-experimental approach, appropriate for two treatments not randomly assigned.
2	Capable of measuring disparity between two groups.	Y	Y	Uses paired data to calculate the Mean Causal Effect of the differences between the two groups (see Eq. 3 in Appendix).
3	Capable of handling a dichotomous dependent variable.	Y	Y	Appropriate for dichotomous or binary dependent variables.
4	Capable of handling multiple independent variables of different types (dichotomous, ordinal, categorical, interval, count).	Y	Y	It can include different types of variables if the assumptions are met.
5	Preserves its power even when there is a significantly smaller sample of one of the observed groups.	Y	Y	Uses paired datasets, reducing issues of lack of representation of a group.
6	Reduces the effect of control variables that are not under analysis.	Y	Y	Uses a matched paired-sample that reduces the effect of control variables.
7	Reduces the number of opportunities to introduce errors and bias from data manipulation.	Y	Y	Sample matching is performed by an algorithm without the analyst fitting the selection criteria and data processing (matching) by trial and error.

Conclusion

Table 7 summarizes the applicability of three statistical methods, for datasets in general and for the SPD dataset empirically tested by OIG.

Table 7. General characteristics for methods[28]	**Propensity Score Matching (PSM)**		**Logistic Regression**		**Logistic Regression with blocked paired-sample**	
OIG evaluation criteria	**General Case**	**Empirical test results**	**General Case**	**Empirical test results**	**General Case**	**Empirical test results**
Observational studies.	Y	Y	Y For quasi-experimental design.	**N**	Y	Y
Measuring disparity between two groups.	Y	Y	Y	Y	Y	Y
Handling a dichotomous dependent variable.	Y	Y	Y	Y	Y	Y
Handling multiple independent variables of different types (dichotomous, ordinal, categorical, interval, count).	Y	Y	Y If both groups are of similar size	**N**	Y	Y
Preserves power even when there is a significantly smaller population of one of the observed groups.	Y	Y	Y If both groups are of similar size	**N**	Y	Y
Reduces effect of control variables that are not under analysis.	Y	Y	**N**	**N**	Y	Y
Reduces number of opportunities to introduce errors and bias from data manipulation.	Y	Y	**N**	**N**	**N**	**N**

[28] The results of the empirical tests for Ordinary Least Squares Regression (OLS) do not appear in this table because the review showed OLS is insufficient to measure effects on disparity

Based on literature and empirical testing, OIG finds that both PSM and logistic regression with blocked paired-sample are appropriate for analyzing the SPD dataset. Specifically, these two methods have the following characteristics:

1) Each is appropriate for observational studies.
2) Each is capable of
   a. measuring disparity between two groups,
   b. handling a dichotomous dependent variable,
   c. handling multiple independent variables of different types (dichotomous, ordinal, categorical, interval, count),
   d. preserving its power even when there is a significantly smaller population of one of the observed groups, and
   e. reducing the effect of control variables that are not under analysis.

PSM outperforms logistic regression with blocked paired-sampling by reducing the number of opportunities to introduce errors and bias due to data manipulation. Particularly when choosing pair-samples, this reduces the possibility of the analyst handpicking which pairs to compare based on the expected outcome of the study.

OIG concludes PSM is an appropriate statistical method to analyze disparity with regard to the race/ethnicity of the subject stopped when compared to various characteristics of the subject, officer or event, for the type and quality of data that SPD has.

Appendix

Frequencies Table

Count and Time Variables

Hour		Weekday		Month		Year	
Min.	0.00	Friday	2041	Min.	1.000	1st Qu.	2016
1st Qu.	5.00	Monday	2299	1st Qu.	3.000	Median	2017
Median	13.00	Saturday	2297	Median	5.000	Mean	2017
Mean	11.91	Sunday	2249	Mean	5.808	3rd Qu.	2017
3rd Qu.	18.00	Thursday	2237	3rd Qu.	8.000	Max.	2018
Max.	23.00	Tuesday	2140	Max.	12.000		
		Wednesday	2200				

Event Description Variables

Priority	
Min.	1.000
1st Qu.	1.000
Median	2.000
Mean	2.576
3rd Qu.	3.000
Max.	9.000

Call Type	
DISPATCH	11228
ONVIEW	4235

Frisk	
No	11880
Yes	3583

Cleared By	
Report Written (No Arrest)	7497
Physical Arrest Made	4785
Assistance Rendered	1620
Unable To Locate Incident Or Complainant	253
Other Report Made	210
Oral Warning Given	190
(Other)	908

Resolution	
-	152
Arrest with GO or Supplemental	4100
Citation / Infraction	59
GO for Prosecutorial Referral	368
GO Report	6722
Street Check	4062

Subject Description Variables

Subject Gender		Subject Age		White/Non-White		Subject Race	
-	11	-	306	1st Qu.	1	White	7837
Female	3234	0 - 17	832	Median	1	Black	4866
Male	12133	18 - 25	3404	Mean	0.833	Hispanic	764
Unable to Determine	85	26 - 35	5250	3rd Qu.	1	Unknown	670
		36 - 45	3135	Max.	1	American Indian / Alaskan Native	542
		46 - 55	1869			Asian	438
		56 and Above	667			(Other)	346

Officer Description Variables

Officer Gender	
F	1713
M	13750

Officer Age	
Min.	21.00
1st Qu.	29.00
Median	33.00
Mean	34.99
3rd Qu.	40.00
Max.	68.00

Years of Experience	
Min.	0.00
1st Qu.	1.00
Median	3.00
Mean	6.26
3rd Qu.	9.00
Max.	37.00

Crisis Interv Trng	
No	7427
Yes	8036

Officer Race	
White	12880
Hispanic or Latino	600
Two or More Races	555
Asian	481
Black or African American	420
Not Specified	301
(Other)	226

Officer Title	
Police Officer	12036
Police Student Officer	1828
Police Officer Probation	1315
Police Sergeant	178
Acting Police Sergeant	57
Acting Police Officer Detective	21
(Other)	28

Squad	
911 Response	13588
ACT SWAT Canine	433
Beats and Bikes	1188
Other	254

Disparity Analysis explained as a random variable

Disparity analysis is a type of analysis where causal effect is a function of potential outcomes, where in terms of random variables, the following is assumed:

Let Y_i be a random variable, where $Y_i(1) \equiv Y_i(T_i = 1)$ is an investigative stop performed when the subject belongs to a non-protected group. And let $Y_i(0) \equiv Y_i(T_i = 0)$ be investigative stop performed when the subject belongs to a protected group. Then the disparity will be a random causal effect described by (Rosenbaum, 2017):

$$Random\ causal\ effect\ for\ unit\ i \equiv Y_i(1) - Y_i(0) \tag{Eq.1}$$

The second causal effect would be Mean Causal effect, described by:

$$Mean\ causal\ effect \equiv E(frandom\ causal\ effect) \tag{Eq.2}$$

$$Mean\ causal\ effect \equiv E[Y_i(1) - Y_i(0)] \tag{Eq.3}$$