November 28, 2017
Casey C. Ross
City of Philadelphia Office of Complete Streets

Philadelphia ARLE 3-0-3 Statistical Analysis – Internal Report

Contents

1. INTRODUCTION

Per the AAA Foundation for Traffic Safety, motor vehicle crashes were the leading cause of death for people aged 16-24 for each year from 2012 through 2014,[i] and 6.3million traffic crashes were reported across the United States in 2015, [up slightly from 6 million in 2014], involving 11.3 million vehicles.[ii] Almost 50% of all vehicles involved in crashes in 2015 were involved in crashes at intersections,[iii] underscoring the importance of safety controls and enforcement at junctions points throughout the roadway system.

As natural points of intermodal interaction and directional conflict, intersections are the sites of most traffic crashes in Philadelphia County: between 2012 and 2016, 32,000 crashes (57% of all crashes in Philadelphia) occurred at an intersection.[iv] During that same period, crashes in which a driver ran a red light accounted for 40% of all reportable crashes in Philadelphia.[v] AAA's 2016 Traffic Safety Culture Index reports that most drivers (82.8% of those surveyed) consider it unacceptable for a driver to run a red light, but more than 1 in 3 drivers (35.5%) admitted to having driving through a light that had just turned red in the past 30 days when they could have stopped safely.[vi] This indicates that in the absence of enforcement, red light running remains an acceptable option for many drivers.

Angle Crashes vs. Rear-End Crashes

Angle crashes are defined as crashes in which vehicles on opposite roadways collide at a point of junction, such as a road intersection, driveway, or entrance ramp. These include "T-bone" crashes as well as right-turn and left-turn crashes. A large body of research undertaken by national policy organizations, think-tanks, academic researchers, and agencies indicates that Angle crashes are some of the most severe, deadly kind of crashes that occur in the United States, especially compared to rear-end collisions.

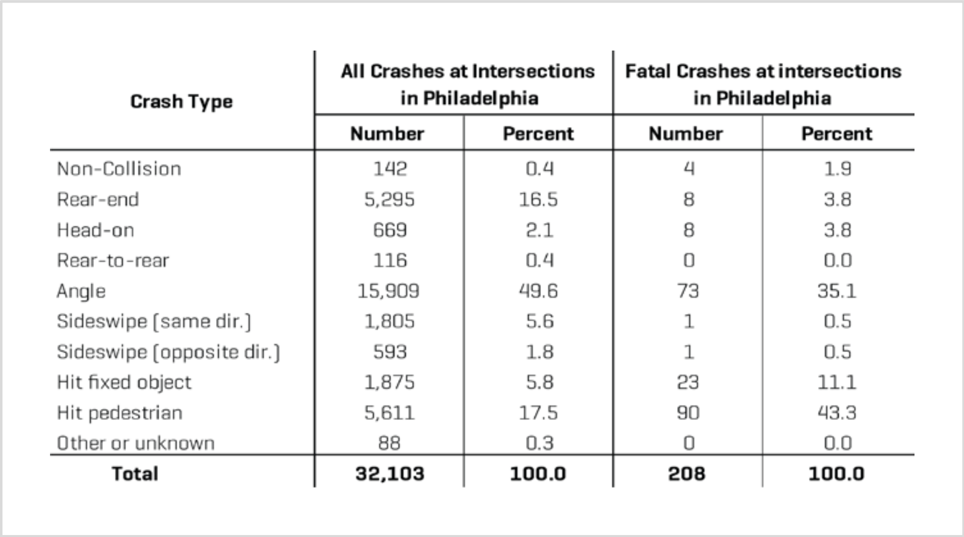
NHTA's 2015 Traffic Safety Facts FARS/GES Annual Report indicates that 14.9% of all fatal Passenger car crashes in 2015 were either left side or right-side angle crashes, while only 1.7% of all fatal crashes were rear collisions. For light trucks (which includes SUVs and pickups), 10.9% of all fatal crashes were right or left side angle crashes, while rear collisions accounted for 1.4% of all fatal crashes. For large trucks (tractor-trailers), 12.8% of fatal crashes were left or right-angle crashes, and 5.0% of crashes were rear-end collisions. For motorcycles, 24.5% of fatal crashes in 2015 were the result of right or left angled collisions, and only 0.4% of all fatal crashes were the result of rear collisions.[vii]

The same report also indicates that in 2015, Angle crashes resulted in more fatalities than any other type of crash involving two vehicles, and resulted in the second-highest number of injuries from crashes. This data set indicates that while rear-end crashes resulted in more property damage and injury than angle crashes, angle crashes were generally more severe: 6,275 vehicle occupants died in left or right-side angle crashes in 2015 compared to 1,480 vehicle occupants killed in rear collisions in the same year.[viii] Another 450,000 vehicle occupants were injured in angle crashes in 2015, compared to 650,000 vehicle occupants injured in rear collisions.[ix]

The most common type crash at Philadelphia intersections are angle crashes, which account for 50% of all crashes (followed by rear-end crashes, which account for 15% of all crashes). Angle crashes are also the second most fatal type of intersection crashes, accounting for 35% of all fatal crashes behind crashes in which a vehicle hit a pedestrian.

From 2012 to 2016, 857 crashes at Philadelphia intersections resulted in a fatality or major injury, accounting for 44% of all crash fatalities in Philadelphia. Broken down by mode, intersection crashes in Philadelphia accounted 53% of all pedestrian fatalities, 41% of bicycle rider fatalities, and 40% of vehicle occupant fatalities. In all, 218 pedestrians, 7 bicycle riders, and 115 vehicle occupants were killed in crashes at Philadelphia intersections between 2012 and 2016. Table 1 provides a detailed breakdown of Intersection Crashes by Crash Type and Crash Severity in Philadelphia from 2012 to 2016.

Table 1: Intersection Crashes by Crash Type and Crash Severity



Crash Type	All Crashes at Intersections in Philadelphia		Fatal Crashes at intersections in Philadelphia	
	Number	Percent	Number	Percent
Non-Collision	142	0.4	4	1.9
Rear-end	5,295	16.5	8	3.8
Head-on	669	2.1	8	3.8
Rear-to-rear	116	0.4	0	0.0
Angle	15,909	49.6	73	35.1
Sideswipe (same dir.)	1,805	5.6	1	0.5
Sideswipe (opposite dir.)	593	1.8	1	0.5
Hit fixed object	1,875	5.8	23	11.1
Hit pedestrian	5,611	17.5	90	43.3
Other or unknown	88	0.3	0	0.0
Total	**32,103**	**100.0**	**208**	**100.0**

Intersections, therefore, are some of the most dangerous places to be a pedestrian in Philadelphia, and are almost as dangerous for vehicle occupants and bicycle riders. Although PennDOT's 2012-2016 data does not include information on whether a crash was related to running a red light, red-light running (RLR) remains a dangerous practice throughout the United States and there is no reason to expect that Philadelphia is in some way an exception.

The Pennsylvania Department of Transportation (PennDOT) compiles information on car crashes in Philadelphia County and makes it publicly available online. This report uses car crash data from Philadelphia County between 2002–2015. ARLE camera enforcement began in Philadelphia beginning in 2005, and new cameras at new locations have been installed as recently as 2015. Per Mike's request, this analysis needed to look at ARLE locations with data for the three years preceding red light camera enforcement and the three years following red light camera enforcement. For this reason, locations installed after 2013 were excluded from the dataset. This resulted in 26 individual locations for analysis:

1. Grant Avenue & Roosevelt Boulevard
2. Red Lion Road & Roosevelt Boulevard
3. Cottman Avenue & Roosevelt Boulevard
4. Broad Street & Oregon Avenue
5. Mascher Street & Roosevelt Boulevard
6. Levick Street & Roosevelt Boulevard
7. Rhawn Street & Roosevelt Boulevard
8. Welsh Road & Roosevelt Boulevard
9. Southampton Road & Roosevelt Boulevard
10. 34th Street & Grays Ferry Avenue
11. 9th Street & Roosevelt Boulevard
12. Broad Street & Hunting Park Avenue
13. 58th Street & Walnut Street
14. JFK Boulevard & Broad Street
15. South Penn Square & Broad Street
16. Aramingo Avenue & Castor Avenue
17. Aramingo Avenue & York Street
18. Henry Avenue & Walnut Lane
19. Rising Sun Avenue & Adams Avenue
20. Broad Street & Vine Street
21. Island Avenue & Lindbergh Boulevard
22. Grant Avenue & Academy Road

23. Bustleton Avenue & Byberry Road
24. Knights Road & Woodhaven Road
25. Knights Road & Woodhaven Road
26. Byberry Road & Worthington Road

Because each intersection is a different size and some are irregular shapes (especially intersections on Roosevelt Boulevard with diagonal crossings), a universal buffer around each intersection would not accurately represent crashes at ARLE locations. Data for each intersection was therefore individually selected from the PennDOT data and coded to its intersection.

Using information on enforcement date provided on page 11 of PennDOT's 2017 ARLE report, I isolated crashes for three years before and after enforcement for each of the 26 ARLE intersections and coded each crash as either pre-ARLE (o) or post-ARLE (1) to create a binary variable for testing.

The final dataset, which consists of crashes at the 26 ARLE intersections for the 3 years preceding and 3 years following ARLE implementation at each intersection, contains 1,244 independent observations, each of which is one crash. Each crash observation has the following variables:

- FID – an ID number assigned to each point in GIS
- CRN – Crash record Number
- CRASH_YEAR – Year in which the crash occurred, ranging from 2002 to 2015
- CRASH_MONTH – Month in which the crash occurred, ranging from 1 to 12
- DAY_OF_WEEK – Day on which the crash occurred, ranging from 1 to 7
- TIME_OF_DAY – Time at which crash occurred, in 24-hour time
- ARLE – a binary variable for which 0 indicates the crash occurred pre-ARLE enforcement and 1 indicates the crash occurred port-ARLE enforcement.
- UUID – An identifier mapping each crash to a specific intersection, ranging from 1 to 26
- FATAL - a binary variable for which 0 indicates the crash resulted in no fatalities 1 indicates the crash resulted in at least one fatality.
- INJURY - a binary variable for which 0 indicates the crash resulted in no injuries of any severity and 1 indicates the crash resulted in at least one injury of any severity.
- MAJ_INJURY - a binary variable for which 0 indicates the crash resulted in no major injuries and 1 indicates the crash resulted in at least one major injury.
- COLL_1 - a binary variable for which 0 indicates the crash was not a rear-end collision and 1 indicates the crash was a rear-end collision.
- COLL_4 - a binary variable for which 0 indicates the crash was not an angled collision and 1 indicates the crash was an angled collision.
- MIN_INJURY - a binary variable for which 0 indicates the crash resulted in no minor injuries and 1 indicates the crash resulted in at least one minor injury.

2. METHODS

2.1 *Problems with OLS Regression if the Dependent Variable is Binary*

OLS regressions works well with a continuous dependent variable (Y), but becomes an ineffective predictive tool if the dependent variable (Y) is binary. A binary Y variable means that the value of Y can be either 0 or 1. Because OLS regression is interpreted as a 1 unit increase in the predictor (x_1) = a β_1 increase in Y, Y can only change from 0 to 1 or 1 to 0 if the dependent variable (Y) is binary. In this situation, the increase in Y by β_1 makes no sense, and OLS regression fails. Logistic regression offers a way to work around this issue by translating OLS regression results into odds, and then transforming those odds into an interpretable result using a log transformation.

2.2 *Solving the Problem with Logistic Regression*

Logistic regression depends on the concept of odds, which allows us to easily interpret the probability of a relationship between the dependent variable and its independent variable(s). In probability, we calculate the number of desirable or undesirable outcomes by dividing by the total number of outcomes. Using a linear method to calculate probability, however, results in numbers that don't make any intuitive sense, including negative probabilities. This is because probabilities need to range between 0 and 1, but linear regression predicts values of Y that range between -∞ and +∞. We use logistic regression to solve this problem.

To calculate odds, we divide the desirable and undesirable outcomes by one another. For example, if there are 100 zip codes and 80 of them have hospitals, the probability of finding a zip code with a hospital is 80/100 = 0.8. The odds of finding a hospital are 80/20 = 4. As the probability increases, the odds increase and vice versa. While probability ranges from 0 to 1, odds range from 0 to +∞.

Mathematically, the odds of an event (Y = 1) can be calculated using the following equation:

$$Odds(Y=1) = \frac{\#\ desirable\ outcomes}{\#\ undesirable\ outcomes} = \frac{\frac{\#\ desirable\ outcomes}{\#\ total\ outcomes}}{\frac{\#\ undesirable\ outcomes}{\#\ total\ outcomes}} \qquad 1$$

$$= \frac{P(Y=1)}{P(Y\neq 1)} = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)} = \frac{p}{1-p}$$

Odds can be transformed to the log of odds through a logarithmic transformation, where the larger the odds, the greater the log of odds. Taking the log of the odds allows us to take a value from - ∞ to + ∞ and return a value from 0 to 1. When the log odds are exponentiated, an odds ratio that ranges from 0 to + ∞ is derived.

The assumptions of the logistic regression are like those of OLS regression, with a few differences. In logistic regression, the dependent variable must be binary. Like OLS regression, there should not be severe multicollinearity. Unlike OLS regression, Logistic Regression requires at least 50 observations per predictor because the regression coefficients are estimated using the maximum likelihood methods (MLE) and not least squares. As in OLS, multicollinearity in Logistic Regression can be calculated using Pearson's correlation between all the predictors. Finally, logistic regression does not assume a linear relationship between the dependent and independent variables, homoscedasticity, or the normality of residuals.

2.3 *Hypothesis for Each Predictor*

Before performing any in-depth analysis, I looked at the correlation between the dependent ARLE variable and the predictor variables: INJURY, FATAL, MAJ_INJURY, UUID, COLL_1, COLL_4, AND MIN_INJURY.

ARLE, INJURY, FATAL, MAJ INJURY, COLL_1, COLL_4, and MIN_INJURY are all binary data sets, while UUID is cardinal. As such, a cross tabulation between the dependent variable and all binary predictors is used to evaluate the binary predictors. Because we have a before and after scenario (pre-ARLE and post-ARLE) paired t-tests are an appropriate method of analysis. For a paired t-test, each dependent variable is split into two categories: pre-ARLE and post-ARLE. These two groups are then compared, and we can either reject each dependent variable's Null Hypothesis or fail to reject each dependent variable's Null Hypothesis:

- **H1 Null:** There is no difference in the average proportion of crashes with fatalities before and after ARLE.
- **H1 Alt:** There is a difference in the average proportion of crashes with fatalities before and after ARLE.

- **H2 Null:** There is no difference in the average proportion of crashes with major injuries before and after ARLE.
- **H2 Alt:** There is a difference in the average proportion of crashes with major injuries before and after ARLE.

- **H3 Null:** There is no difference in the average proportion of crashes with minor injuries before and after ARLE.
- **H3 Alt:** There is a difference in the average proportion of crashes with minor injuries before and after ARLE.

- **H4 Null:** There is no difference in the average proportion of rear-end crashes before and after ARLE.
- **H4 Alt:** There is a difference in the average proportion of rear-end crashes before and after ARLE.

- **H5 Null:** There is no difference in the average proportion of angle crashes before and after ARLE.
- **H5 Alt:** There is a difference in the average proportion of angle crashes before and after ARLE.

If the p-value is less than 0.05, we reject the null hypothesis for the alternative. For this analysis, each paired t-test was performed twice: one on dependent variables as raw data, and once on dependent variables as proportions of all crashes.

3. RESULTS

3.1 Paired t-Test Results

The results for all paired t-tests are shown in Tables 2 and 3 below:

Table 2: Paired t-test results with dependent variables as raw data

HYPOTHESIS	VARIABLE	T-STATISTIC	P-VALUE	OUTCOME
H1	No Fatal Injury	0	0.5	Fail to reject the Null Hypothesis
	Fatal Injury			
H2	No Major Injury	-0.25351	0.599	Fail to reject the Null Hypothesis
	Major Injury			
H3	No Minor Injury	1.7489	0.04628	**Reject the Null Hypothesis in favour of the Alternative Hypothesis**
	Minor Injury			
H4	Rear-End Crash	-0.72509	0.7624	Fail to reject the Null Hypothesis
	Other Crash			
H5	Angle Crash	0.1853	0.4273	Fail to reject the Null Hypothesis
	Other Crash			

Table 3: Paired t-test results with dependent variables as proportions of all crashes

HYPOTHESIS	VARIABLE	T-STATISTIC	P-VALUE	OUTCOME
H1	No Fatal Injury	-0.21126	0.5828	Fail to reject the Null Hypothesis
	Fatal Injury			
H2	No Major Injury	-0.78588	0.7803	Fail to reject the Null Hypothesis
	Major Injury			
H3	No Minor Injury	2.0126	0.02752	**Reject the Null Hypothesis in favour of the Alternative Hypothesis**
	Minor Injury			
H4	Rear-End Crash	-1.3069	0.8984	Fail to reject the Null Hypothesis
	Other Crash			
H5	Angle Crash	-0.66737	0.7447	Fail to reject the Null Hypothesis
	Other Crash			

For H1, H2, H4, and H5 we fail to reject the null hypothesis and conclude that at these 26 intersections, the proportion of crashes involving fatalities, major injuries a rear-end collision, and an angle collision are not significantly different pre- and post-ARLE.

For H3, we reject the Null Hypothesis in favour of the Alternate Hypothesis and conclude that the proportion of crashes involving a minor injury are significantly different pre- and post-ARLE.

3.2 *Regression Assumption Checks*

Unlike OLS regression, Logistic regression does not assume a linear relationship between the dependent variable and predictor variables, does not assume homoscedasticity, and does not assume normality of residuals. As such, no tests for these assumptions are required.

Logistic Regression assumes that the Dependent variable is binary, and that assumption is met in this case. Logistic regression also assumes that you have 50 observations per predictor. This model has 6 predictor variables, which means the data must have at least 300 observations. This data set has 1,244 observations, so that assumption is also met. Finally, Logistic Regression also assumes that there is no severe multicollinearity. Testing for multicollinearity requires looking at Pearson correlation for the dependent variable and each predictor variable.

Pearson correlation measures the strength of the linear relationship between two variables. The value of the results of Pearson correlation always fall between -1 and +1, where -1 indicates a perfect negative relationship between the two variables, +1 indicates a perfect positive relationship between the two variables, and 0 indicates no linear relationship between the two variables (slope = 0).

- When 0 < R < 0.5, the variables x and y have a weak linear relationship.
- when 0.5 < R < 0.8, the variables x and y have a medium-strength linear relationship.
- When 0.8 < R < 1, the variables x and y have a strong linear relationship.

The results of the Pearson correlation cross-table test for the predictor variables are shown in Table 5 below:

Table 4: Pearson Cross-Tabulation correlation table for predictor variables

	FATAL	**MAJ_INJURY**	**COLL_1**	**COLL_4**	**MIN_INJURY**
FATAL	-	-0.0112	-0.0409	-0.0376	-0.0455
MAJOR INJURY	-0.0112	-	-0.0551	-0.0093	-0.0471
COLL_1	-0.0409	-0.0551	-	-0.5535	-0.5535
COLL_4	-0.0376	-0.0093	-0.5535	-	-0.0492
MIN_INJURY	-0.0455	-0.0471	0.0754	-0.0492	-

These results indicate that there is no severe multicollinearity present in the predictor variables, which means they can all be used in our regression model without violating any assumptions.

Because most our predictors are binary, Pearson correlation is not the best means of testing for collinearity. For a more rigorous statistical analysis, looking at the mean square contingency coefficient (the Phi coefficient) would be a better test.

3.3 Logistic Regression Results

The table below shows the results of a logistic regression model that includes all predictor variables, binary and continuous, regardless of their Chi-Square performance:

Table 5: Results of Logistic Regression Model with all predictors

	Estimate	Std. Error	z value	Pr(>|z|)	Sig Code	OR	2.50%	97.50%
(Intercept)	0.03855252	0.1606373	0.23999733	0.81033233	-	1.0393053	0.7586179	1.42494
FATAL	0.05089141	0.6423282	0.07922961	0.93685	-	1.0522086	0.287403	3.849808
MAJ_INJURY	0.04417151	0.4679884	0.09438592	0.9248026	-	1.0451616	0.4087444	2.63532
COLL_1	0.34507398	0.1495474	2.30745634	0.02102939	*	1.4120944	1.0539302	1.894608
COLL_4	0.19033589	0.1404683	1.35500907	0.17541467	-	1.2096558	0.9188508	1.593996
MIN_INJURY	-0.20897219	0.1281862	-1.63022351	0.10305427	-	0.8114178	0.63094	1.043014

These results indicate that only the COLL_1 predictor variable is significantly related to the dependent variable ARLE. The significance is minimal.

In this model, we can say that a 1 unit increase in a predictor variable corresponds to a ($e^{\beta 1}$-1) * 100% change in the odds of Y=1, holding the values of the other predictors constant. We can interpret the outcome for each significant predictor variable as follows:

- The odds of a car crash occurring post-ARLE go up by ($e^{\beta 1}$-1) * 100% = ($e^{0.314004}$ - 1) * 100% = 37% if the crash was collision type 1 (a rear-end collision), holding all other variables constant.

4. CONCLUSIONS & NEXT STEPS

This data set presents several challenges for testing:

1. In some cases, as in the case of fatal crashes and crashes with major injuries, the data set has small values and is irregularly distributed.

2. There is no control group to which ARLE intersections can be compared, so the actual change has no true point of comparison.

Despite these challenges, this analysis does allow us to make some conclusions.

The results of the Paired t-Tests are generally inconclusive. H3 has a significant result, indicating that there is a statistically significant difference in minor injuries between pre-AREL and post-ARLE, and that the difference is positive post-ARLE. This fits the hypothesis that ARLE implementation may result in an increase in less-severe crashes. Unfortunately, because H1 and H2 are inconclusive, we do not have a full picture confirming this hypothesised trend. H4 and H5 have p-values close to 1 and therefore cannot be interpreted meaningfully.

The logistic regression results indicate one important thing: crashes occurring post ARLE implementation are more likely to be rear-end collisions. These findings track with the expectation that ARLE implementation might result in a higher overall incidence of crashes, especially rear-end crashes, but that the severity of crashes will be lessened.

Because none of the other predictor variables were significantly related to the dependent variable, drawing conclusions from their results in Table 4 would be questionable. In the case of the FATAL variable and the MAJ_INJURY variable, the issue is likely that the data sets are so small: as Table 2 showed, only 10 crashes total resulted I fatalities at all ARLE locations between 2002 and 2015, and only 19 resulted in a major injury.

It is surprising that the COLL_4 variable did not have any significant relationship to the dependent variable: one of the central arguments for ARLE implementation is the program's potential to decrease the rate of angle crashes at high-volume and dangerous intersections. In the case of the present data set, ARLE implementation and Angle crashes do not appear to have a statistically significant relationship.

It might be worthwhile to look at data for 5 years pre- and post-ARLE in a future analysis to try and get a larger data set. Because fewer intersections have 10 years of data to look at, however, the data set might not change dramatically in size, and the data would be (generally) older.

It would also be appropriate to find intersections that technically meet the criteria for ARLE but do not have any red-light cameras and use them as a control group, like the analysis undertaken in the PennDOT Pennsylvania Automated Red-Light Enforcement 2017 Program Evaluation.

[i] AAA Foundation for Traffic Safety, "2016 Traffic Safety Culture Index" (Washington DC: AAA Foundation for Traffic Safety, February 2017), 1, https://www.aaafoundation.org/sites/default/files/2016TrafficSafetyCultureIndexReportandCover_0.pdf.

[ii] US Department of Transportation National Highway Traffic Safety Administration, "TRAFFIC SAFETY FACTS 2015," 78.

[iii] Ibid.

[iv] "PennDOT Crash Download Map" (Harrisburg, PA: Pennsylvania Department of Transportation), accessed September 20, 2017, https://pennshare.maps.arcgis.com/apps/webappviewer/index.html?id=8fdbf046e36e41649bbfd9d7dd7c7e7e.

[v] "PennDOT Reportable Crash Statistics," Technical (Harrisburg, PA: Pennsylvania Department of Transportation, 2017), https://www.dotcrashinfo.pa.gov/PCIT/welcome.html.

[vi] AAA Foundation for Traffic Safety, "2016 Traffic Safety Culture Index," 5.

[vii] US Department of Transportation National Highway Traffic Safety Administration, "TRAFFIC SAFETY FACTS 2015," 92–100.

[viii] Ibid., 70.

[ix] Ibid., 125.