Botts' Dots Effectiveness

Report on Raised Pavement Markers and Reducing Traffic Accidents

Jacqueline Zawada

May 15, 2020

Contents:

Abstract

Botts' dots are a tool used on roadways to limit traffic and car accidents. The following report studies the effectiveness of these dots in reducing traffic accidents. The data used for this investigation comes from the 1000 different road segments in California, in which data was collected each year starting in 1980 and ending in 1989. Exploratory analysis shows that the average number of accidents was lower in locations where the dots were installed as compared to locations where they were not installed. However, this does not control for other variables. In order to take into account other road conditions two regression methods were considered: Poisson and negative binomial. A comparison of these models showed that the negative binomial model was the best. The results from this model tell us that Botts' dots do in fact lower the number of crashes that occur. Next, in order to take into account the effect that time has, we added a variable to the model that accounts for time. This showed that accidents decreased with time, and this added control decreased the overall effect that Botts' dots had on reducing accidents.

1. Introduction

Raised pavement markers, otherwise known as Bott's dots, have been used since the 1980's to try and mitigate traffic related issues. They are raised discs that get glued on the center line of roads, with the intention to reduce speeding and traffic accidents. While Botts' dots are relatively cheap and easy to install, they are more work to maintain than alternative options like speedbumps. This is why it is important to ensure that Bott's dots are successful their aim to reducing traffic accidents. The following report will analyze car crash data from the state of California, and investigate the effectiveness of these pavement markers in limiting crashes.

2. Exploratory Analysis

Data

The dataset was collected by gathering the number of traffic accidents and road information from 1,000 different road segments within California. The data was collected for the duration of the 1980's, which resulted in 10,000 total observations, and no missing data. In addition to recording the number of crashes that occurred at that given location, there is also information about the road, such posted speed limit and number of intersections. Each location also has and indicator variable that shows whether or no Bott's dots were installed.

The following list shows which variables are provided in this data set:

- *crash*: total number of crashes
- *Botts*: indicator of whether Botts' dot was installed (1: yes, 0: no)
- *adt*: annual average daily traffic volume (unit is vehicles per day)
- *length*: roadway segment length in miles
- *width*: three categories of pavement width in feet:
  - Three levels: less than 20, 20 – 24, more than 24
- *speed*: posted speed limit

    - levels: less than or equal to 55 mph, greater than 55 mph
- *shoulder*: three categories of average shoulder width in feet
    - levels: less than or equal to 3, 3 – 6, greater than or equal to 6
- *driveways*: three categories of number of driveways
    - levels: none, 1 – 10, greater than or equal to 11
- *intersections*: two categories of intersections:
    - levels: no intersection, at least 1 intersection
- *curves*: existence of horizontal curves:
    - levels: no curves, at least 1 curve
- *curvature*: average degree of curvature
- *year*: the calendar year in which the road site information was recorded
- *location*: the road site index

Summary Statistics:
As the list above shows, there are some variables, which are represented categorically, and some that are represented numerically. Table 1 shows the summary statistics of 4 of the variables represented numerically. Note that the variable *crash* is a count of the number of total number of accidents, thus it will only take on integer values within the dataset.

adt	length	curvature	crash
Min. : 351.4	Min. :0.0230	Min. : 0.000	Min. : 0.000
1st Qu.: 2254.8	1st Qu.:0.4060	1st Qu.: 0.000	1st Qu.: 0.000
Median : 3594.6	Median :0.4770	Median : 1.996	Median : 1.000
Mean : 4590.7	Mean :0.4766	Mean : 3.511	Mean : 1.556
3rd Qu.: 6293.3	3rd Qu.:0.5703	3rd Qu.: 4.823	3rd Qu.: 2.000
Max. :21785.4	Max. :0.7570	Max. :77.829	Max. :18.000

Table 1: Summary of continuous variables

We can gather important information from Table 1. The variable adt represents the annual average daily traffic volume. There appears to be a fairly large spread to this data, with the smallest value being 351, and the largest being 21,785. This will likely be an important variable to take into consideration because the amount of traffic can have a great effect on how many crashes occur. Curvature, which measure the average degree of curvature, suggest that there are a decent number of roads that don't have any curves. The minimum and the firs quartile are both zero, which means that at least 25% of the 1000 locations did not have any curves. In addition, the maximum value is 77.8, which is much higher than the 3rd Quartile value. This suggests that there are a few locations that have way sharper curves, compared to the rest of the locations. The mean for the number of crashes is 1.556, with a median of 1. Noticeably, the max number of crashes is 18, which is significantly higher than those measures of center. This suggests that there is positive skew, or right skew, to the distribution. This distribution can be seen in Figure 1 in the next section.

We suspect that there is a relationship between areas that have more traffic and how wide the road is. Generally, higher trafficked roads tend have more lanes and thus would be wider.

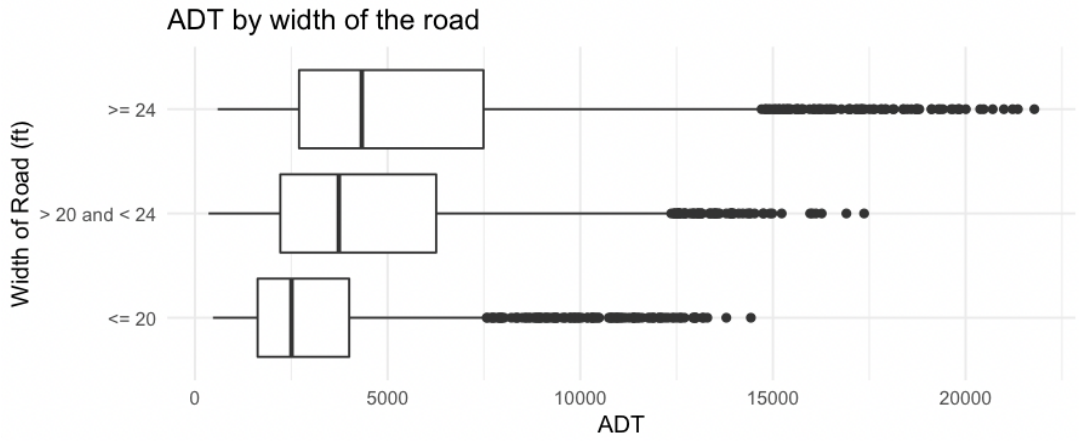

Figure 1: Annual average daily traffic by Width of the road

The plot above shows the annual average daily traffic volume separated by how wide the road is. It shows that the amount of traffic does tend to be higher on roads that are wider.

Investigation into Number of Crashes:

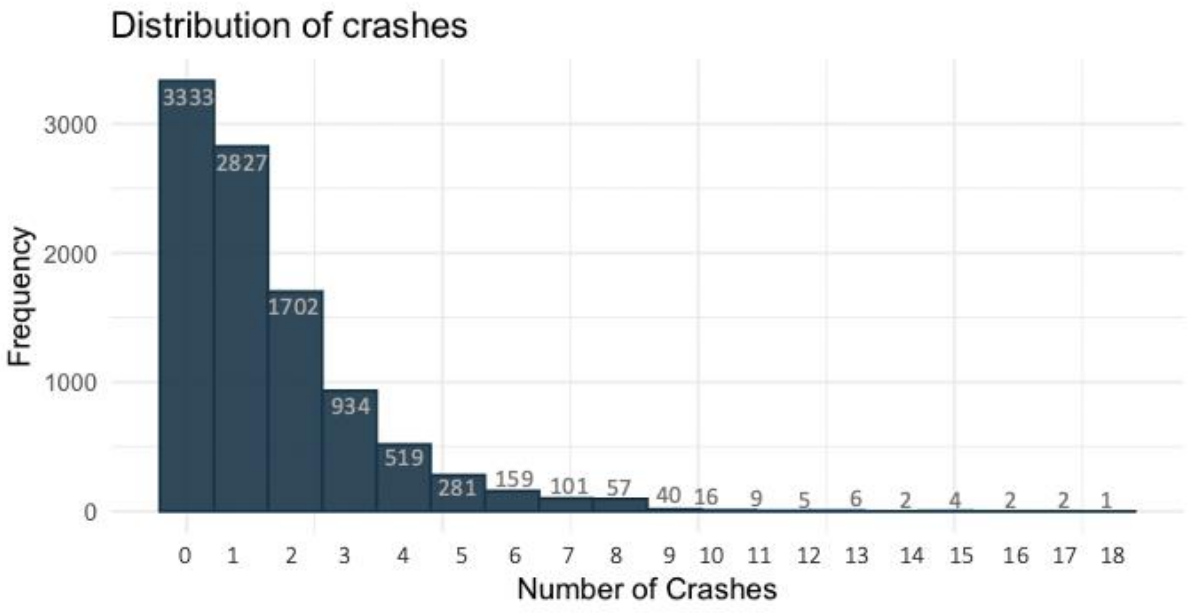

Figure 2: Histogram of number of crashes

Figure 2 shows the distribution of the number of crashes. We can see an clear skew in this data, with a majority of the observations having 2 crashes or fewer. Furthermore, there were only 84 observations, out of the total

10,000 that had 9 or more crashes occur. It is possible that there is one or two road segments in the dataset that is responsible for producing more crashes. For example, if a road has a sharp curve or a lot of traffic, more accidents may occur there.

Investigating this further showed that for the most part these high number of crashes happened somewhat randomly across locations. However, there were two locations that did produce noticeably more crashes. Location 125 had 10 occurrences of 9 or more crashes out of the 84 total, and location 765 had 8 occurrences. Table 2 shows the information about these two road segments. Both locations utilized Bott's dots, with location 125 introducing them 1 year in, and location 765 introducing them 7 years in. Location 125 is a wide road with low speed and no curves, which should all be conducive to lower crashes. The notable statistic for this section is the average annual daily traffic is 11,376, which is much larger compared to other road segments. Referring back to Table 1, the mean ADT is 4,590.7. Location 765 has a curvature of 5.05, which (referring to Table 1) is slightly over the 3 quartile, meaning that 75% of the roads have a lower curvature. This segment of road also has high speed, which could cause more crashes, especially on a more curved road. In addition, the mean ADT for this road segment is 17, 946, which is even higher than that of location 125. The combination of these factors could help explain why so many crashes occur on this segment of the road.

	Location 125	**Location 765**
Botts' Dots	Introduced in 1981	Introduced in 1987
Mean ADT	11,376.6	17,946.72
Length	0.426	0.467
Width	≥ 24	≥ 24
Speed	Low	High
Shoulder	> 6	> 6
Driveways	> 0 and ≤ 10	> 10
Intersections	None	≥ 1
Curvature	0	5.05
Mean crashes	12.6	12.3

Table 2: Details about road segments 125 and 765, which produced high numbers of crash

Investigation into Bott's Dots:

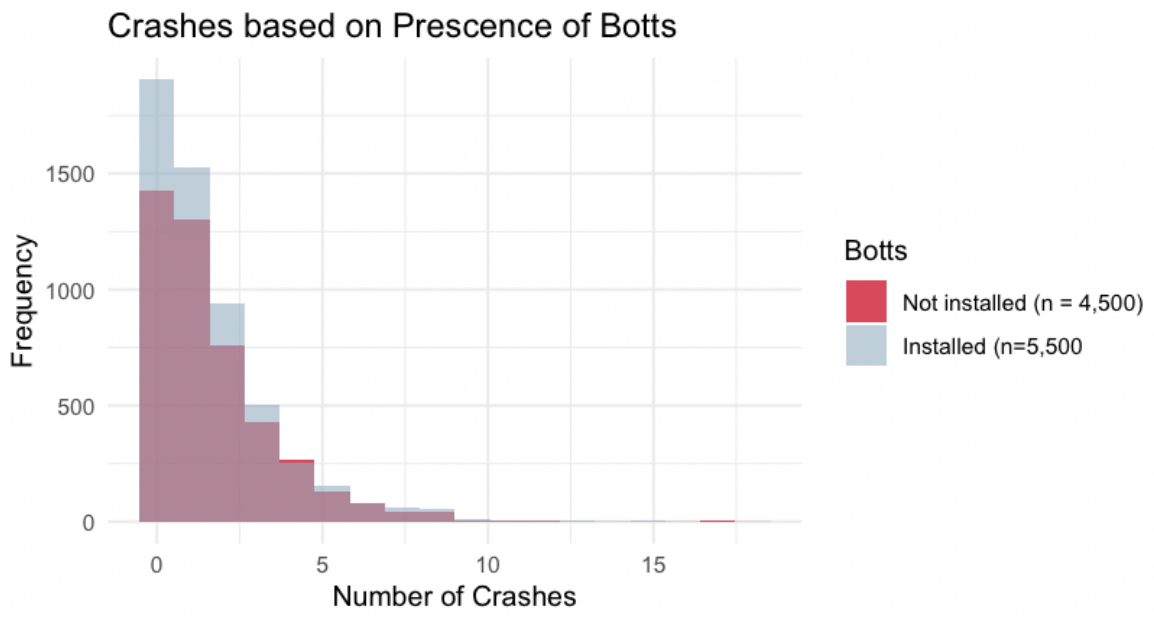

Figure 3: Histogram of number of crashes based on whether or not Botts' Dots were installed

The data set has more observations where Botts' dots where installed, which is reflected in the histogram in Figure 3. In general, the distribution of crashes with Botts' dots appears to be roughly the same as the distribution of crashes with the dots installed. It is hard to tell from the histogram if the dots have any significant relationship to the number of car accidents.

The mean number of crashes when the dots are present is 1.521, and the mean number of crashes when they are not present in 1.600. These two means seem to be pretty close to each other, however, a majority of the data falls below 2, so minor differences in the means are to be expected here. A two sample t-test allows us to compare these means and test if the mean number of crashes where the dots are installed is in fact significantly lower. This hypothesis test does not take into account the other variables about the road, which is important when making conclusions about the overall effectiveness of the dots. However, it will help us get a basic idea of the relationship between the dots and whether or not they reduce accidents. The hypothesis for this test will be:

$$H_0: \mu_{with\ Botts} < \mu_{without\ Botts}$$
$$H_1: \mu_{with\ Botts} \geq \mu_{without\ Botts}$$

This one-sided two sample t-test gives a test statistic of -2.11 on 9,998 degrees of freedom, which corresponds to a p-value of 0.017. Thus at the level $\alpha = 0.05$, the mean number of crashes with Dotts present is significantly lower.

3. Regression Analysis

While the t-test showed us the average number of crashes is lower when Botts' dots are installed, this does not lead to the conclusion that they are effective. As discussed earlier, there are multiple factors that could impact how many crashes there are, such as how much traffic there is, how many intersections there are, and if there are any sharp turns. Without controlling for the effect of these variables, we cannot make a conclusion on the effectiveness of Botts' dots.

Regression Analysis Methods

The dependent variable in this analysis is the number of car crashes that occurred, which is a count.
As the histogram in Figure 2 shows, the distribution of crashes is highly skewed. For these reasons, it would not be appropriate to do an OLS regression on this data. The crash variable can only take on integer values, and even with a data transformation, the distribution does not resemble a normal curve. Fortunately, there are models that are more appropriate to use in situations where the dependent variable is a count, and has a highly skewed distribution.

I. **Poisson:** One option for regression models that we considered is a Poisson model. Poisson regression models are useful when you want to model count data which follows a Poisson distribution. The Poisson distribution models the probability of y events occurring, and has the property that the mean and variance are equal. This is an assumption that should be met when using Poisson regression. The pdf for a Poisson distribution is:

$$P(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

We can see that our data follows a similar shape to a Poisson distribution with a rate parameter of $\lambda = 1$.

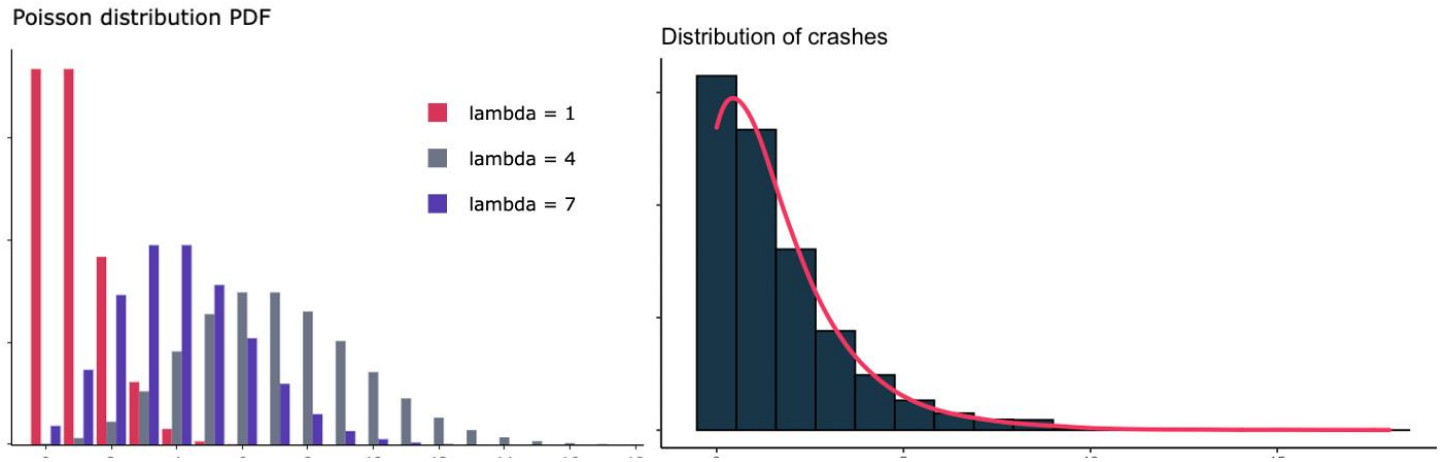

Figure 4: The plot on the left shows the distribution of randomly generated Poisson variables for 3 different rate values. the plot on the left shows the distribution of crashes.

Poisson regression transforms the non-linear relationship between the dependent variables and the predictor into a linear one using a log link function. Thus, the general mathematical form of this model is:

$$\log(Y_i) = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k$$

$$Y_i = e^{\beta_0 + \beta_1 x_1 + \cdots \beta_k x_k}$$

II. **Negative Binomial:** The second method we considered is negative binomial regression. Negative binomial regression is useful when there is overdispersion in the count data. Overdispersion occurs when the conditional variance of the count variable exceeds the condition mean, which violates the assumption that Poisson regression uses. In the data used for this report, the average number of crashes is 1.1556, with a variance of 3.385, which suggests that there may be an overdispersion issue with the Poisson model. Negative binomial regression is considered to be a generalization of the Poisson regression. The model will have the same mean structure as the Poisson model, however it has an extra parameter which models the overdispersion. The negative binomial model generally has coefficient estimates that are similar to that of the respective Poisson model, however the standard errors for these Negative Binomial estimates will likely be larger.

Model Selection

Before creating the regression models, we first considered what predictor variables may be useful as controls. It is reasonable to assume that most of the variables provided in the dataset would be valuable to control for. However, we did choose to exclude some of them. The location index was not included, because most of the variation in location is already captured with the other variables. The curve and curvature variables capture essentially the same information, however curvature gives a more precise estimate, so we chose to use this measure over the categorical curve variable. Lastly, at this point we are assuming the road conditions do not change with time, so the year variable was excluded from consideration.

For both regression methods, we started by fitting a full model using the following variables as predictors: botts, adt, length, width, speed, shoulder, driveways, intersections, curvature. We used the likelihood ratio statistic to compare nested models, until a satisfactory final model was achieved. The likelihood ratio is calculated by finding the difference of the likelihoods of the nested models we wish to compare, and then multiplying this difference by 2. This statistic follows a chi-squared distribution. If this statistic is found to be significant, then we can conclude that the variables removed to create the smaller model did not contribute significantly to the model, and thus the smaller model is more appropriate.

Both Poisson regression and negative binomial regression ended up using the same variables in the final model. These variables were: botts, adt, width, shoulder, driveways, intersections, and curvature. So, length and speed were determined to not be significant. Based on the exploratory analysis, we saw there may be a relationship between the average annual daily traffic, and the width of the roads, so we also included a model that included the interaction of these two variables.

Comparison of Models

	Poisson	Negative Binomial	Negative Binomial with Interaction
AIC	31,716	30, 918	30, 892
Residual Deviance	14,568 df: 9,989	10,721 df: 9,989	10,722 df: 9987
Null Deviance	19,901 df: 9,999	14,486 df: 9,999	14,530 df: 9999

Table 3: Comparison of final Poisson model and final negative binomial model.

In comparing the three models, we see that the AIC is lower for the negative binomial regression models, which indicates one of these may be the better model. This aligns with what was expected due to our concerns about overdispersion. We can check for overdispersion in the Poisson model by comparing the residual deviance to the degrees of freedom. If the residual deviance is larger than the degrees of freedom, than there is overdispersion. From Table 3 we see this is the case. The Poisson model has a residual deviance of 14,568 and only 9,989 degrees of freedom. The fact that we have overdispersion, tells us that the mean number of crashes is less than the variance, which violates a key assumption in using Poisson regression. For these reasons, we will use the Negative Binomial model to proceed. The AIC of the model with the interaction term is the lowest out of the three, and the liklihood ratio test showed that this model was better compared to the Negative Binomial model without the interaction term, thus we will use this model to estimate the effect of Botts' dots.

Results and Interpretation

Negative Binomial Regression Model Estimates

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6405	0.05126	-12.49	8.065e-36
botts	-0.1796	0.01993	-9.011	2.049e-19
adt	0.0002071	8.629e-06	24	2.888e-127
width> 20 and < 24	0.3724	0.05431	6.858	6.991e-12
width>= 24	0.4042	0.05623	7.189	6.517e-13
shoulder> 3 and <= 6	0.03119	0.02197	1.419	0.1558
shoulder> 6	-0.05635	0.03635	-1.551	0.121
driveways> 10	0.1002	0.02247	4.462	8.134e-06
drivewaysNone	-0.05459	0.04767	-1.145	0.2521
intersectionsNone	-0.07764	0.02224	-3.492	0.00048
curvature	-0.002593	0.001831	-1.416	0.1568
adt:width> 20 and < 24	-5.087e-05	9.722e-06	-5.232	1.676e-07
adt:width>= 24	-5.49e-05	9.537e-06	-5.756	8.59e-09

Table 4: Coefficient estimates for the final Negative Binomial regression model

Table 4 shows the results of the negative binomial regression model with the interaction term. The estimate associated with botts will give insight into how effective Botts' dots are. The value of the botts coefficient

estimate is -0.18. This estimate has a p- value that is very low, which tells us that it is significantly different than zero.
To interpret this effect of botts, we will need to calculate $e^{-0.18} = 0.834$. The reason we do this is because we had to use the log link function to create the model. Figure 5 graphs the values of all the estimates raised to the power of e, and shows their corresponding 95% confidence interval. This estimate for Botts tells us that the locations where Botts' dots are installed had 0.83 times fewer accidents. Or, in other words, the presence of Botts' dots is associated with a 17% reduction in car crashes. Overall, we observe that when we control for other road conditions, Botts' dots do in fact help to reduce the number of car crashes that occur.

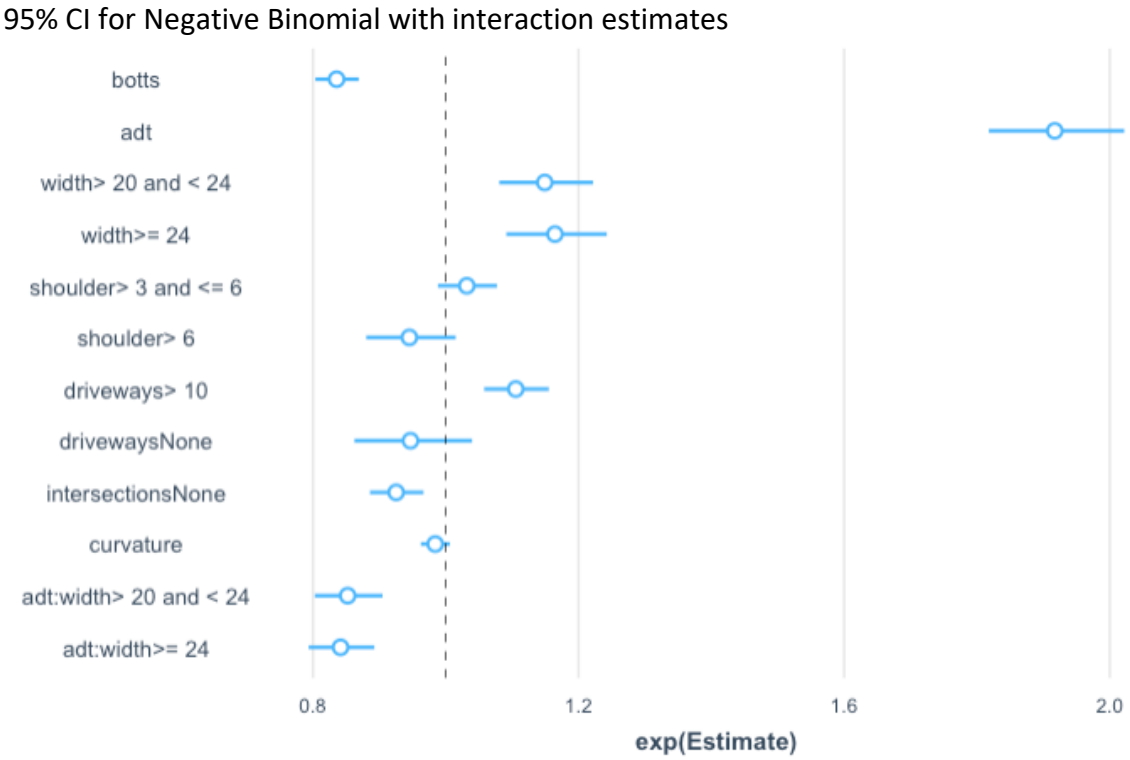

Figure 5: This graph shows the 95% confidence intervals of the estimates from the Negative Binomial model. The estimates have been raised to the power of e.

4. Trend over time

Up to this point, we have assumed that conditions have stayed constant over time. However, it is possible that the number of crashes is affected by time. To get a basic idea of how crashes are related to the year, we can take the average number of crashes for all 1,000 locations for each of the 10 years that data was gathered from 1980 to 1981.

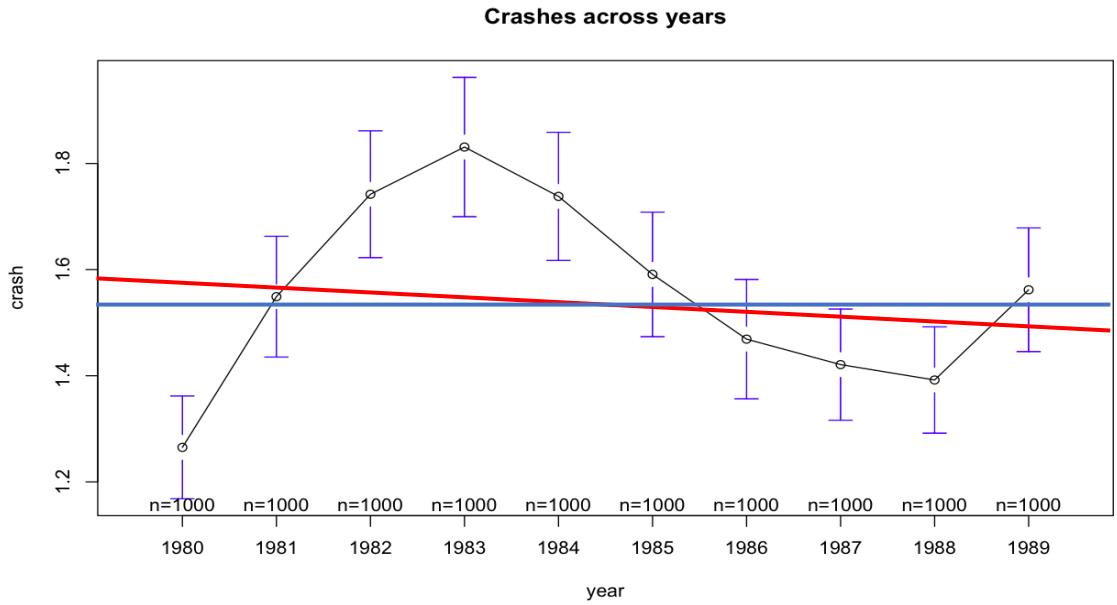

Figure 6: Average number of crashes each year, with the red line indicating the trend

Figure 6 shows the plot of the average crashes over all 1,000 road locations by year. The blue line indicated what no trend would be, and the red line indicated the trend of the data. This plot suggest that the number of crashes does appear to decrease some as the year continue. This trend does not appear to be very extreme, however, the average number of crashes does not have a very large range, so even a slight trend can be meaningful.

We account for the time by adding a variable to the negative binomial regression model. To do this we will create a variable that is an integer representing how many years it has been since 1980. To start, let's look at a model that just uses that time variable as a predictor, and the number of crashes as the dependent variable.

Coefficient Estimates for negative binomial model, using only years since 1980 as a predictor.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.4648	0.0212	21.93	1.379e-106
I(year - 1980)	-0.005066	0.003983	-1.272	0.2035

These results show that there is a negative relationship between the number of crashes and the year. However, the estimate is not significantly different than 0, seeing as it has a p-value of 0.20. In order to understand the true effect that time has, we will need to control for the other road conditions that we have found to be relevant to estimating the number of crashes. We will use the final model created in the previous section, however now we will add in the variable that represents the years it has been since 1980.

In this final model the variables we are including are: botts, adt, width, shoulder, driveways, intersections, curvature, an interaction between adt and width, and the number of years since 1980. Note that the table below does not show the estimates for all of the variables included in the model.

Coefficient Estimates of Relevant variables for final model, controlling for time

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.5504	0.05176	-10.63	2.07e-26
botts	-0.05535	0.02406	-2.3	0.02143
I(year - 1980)	-0.03893	0.004268	-9.122	7.39e-20

This table shows the estimates for only two of the variables used in the Negative Binomial Regression

The estimate for botts and the estimate for time are both statistically significant in this larger model. The estimate for years since 1980 is negative, which confirms the theory that the number of crashes has reduced with time. We can interpret this estimate by calculating $e^{-0.04} = 0.96$. This means that each additional year results in .96 times fewer accidents. Or, an increase in one year decreases the number of accidents by 4%. The coefficient estimate for botts has decrease in this model, which suggests that the time variable is controlling for some of the changes in accident numbers, that wasn't reflected in the previous model. With this new model, we conclude that the presence of Botts' dots is associated with a 5% reduction in car crashes.

This model suggests that there is a negative trend between time and the number of car crashes. And when we control for this trend, we find that the effect of Botts' dots in reducing accidents decreases quite a bit, for reducing accidents by 17% to 4%.

5. Conclusion

This report aimed to understand how effective Botts' dots were at reducing traffic accidents. We began by doing a time two- sample T-test to compare the mean number of crashes at locations where they were installed to those where they were not installed. What we found is that the average number of crashes is lower in locations where Botts' dots are installed. However, as was discussed, this test does not allow us to control for road conditions that could also affect the number of crashes. Thus we proceeded to find a regression model that would allow us to control for the pertinent road conditions. We found the a negative binomial model was best suited for the data, and created a model that controlled for annual average daily traffic volume, width of the road, average shoulder width, number of driveways, number of intersections, and degree of curvature. The model also includes an interaction between average annual traffic and width of road. The results of this model showed that Botts' dots do have a significant effect on the number of crashes. Specifically, we found that the locations with Botts' dots installed resulted in a 17% decrease in the number of car crashes. Lastly, we considered the possibility that time may play a role in affecting how many crashes there are. We saw that there appeared to be a general negative relationship between the year and the number of crashes. This was confirmed when we added a variable to control for time into the model. We saw that this variable did show a negative trend between time and car crashes. Adding this variable also greatly reduced the effect that Botts' dots had on reducing car crashes. Overall, we have found that the is some relationship between the installation of Botts' dots and reducing accidents.

However, when we control for other variables, such as road conditions and time, we find that the effect that they have is relatively small.

Variables included	Chi-squared and p-value
Botts, adt, width, speed, shoulder, driveways, intersection, curvature, length	
Removed: length	$\chi^2_1 = 0.011$, p = 0.92
Removed: speed	$\chi^2_1 = 2.46$, p = 0.12
Removed: driveways	$\chi^2_1 = 35.69$, p = 1.78e-8
