How does weather affect the amount of traffic accidents in California from 2010 to 2017?

A short report on non-parametric regression for Advanced Econometrics 2

Dante van der Heijden (11020075), Willem Kullberg (11041544) and Wietse Steenstra (11004487)

January 22, 2021

Lecturer: *Dhr. dr. K.J. (Kees Jan) van Garderen*

Abstract
In this paper we explore the relationship between temperature and frequency of traffic incidents, comparing the results of a parametric and non-parametric regression approach. Temperature and daily collisions are found to be slightly positively correlated. Although both regression methods show similar predictions, the confidence interval of the non-parametric regression estimates is more in line with the actual data, making this the preferred method.

1 Introduction

In the motored world of today, traffic accidents have become an expected risk of travelling and are part of everyday life. Nonetheless, its impact on society remains ever so large, with car crashes causing over 38,000 yearly fatalities in the U.S. alone[1]. Numerous government departments have been founded to mitigate the impact of these incidents, providing help to victims and ensuring that roads remain as safe as possible.

For these institutes to allocate their resources (like towing trucks or road-side personnel) efficiently, they need accurate predictions of accident frequency on any given day. According to Zou, Zhang, and Cheng (2021), the climate and weather circumstances can be important predictors for this frequency. Leard, Roth, et al. (2015) find the same thing and highlight temperature as their most important variable. Building on their work, we investigate the effect of temperature on the amount of traffic incidents in the state of California, using a non-parametric approach. To investigate whether the unrestrained functional form can improve our results, we perform a standard OLS and a kernel density regression, after which we compare the results. The rest of this document is structured as follows: Section 2 describes our methodology, the results are presented in Section 3 and Section four concludes.

2 Methodology

2.1 Dataset

We retrieve public traffic collision data from 2010 to 2017 for the state of California from the Statewide Integrated Traffic Records System (SWITRS). For each collision we match the location to the closest weather station and use NASA's API[2] to obtain the historical climate data. It is assumed that precipitation will be an important covariate and that the distribution of accidents will change when we condition on this. However, since our dataset has relatively little observations with rain or snow present, we decide to filter out these entries and only include accidents that occurred during dry weather. Finally, we group this data set per day to find the total number of accidents and the average temperature at which these accidents happened. In Figure 1 we see that this variable is skewed left and

[1] U.S. Department of Transportation's Fatality Analysis Reporting System (FARS)

[2] https://power.larc.nasa.gov/docs/services/api/

Figure 2 shows an increase over time.

2.2 Parametric Regression

For our benchmark parametric model, we choose a simple fourth-order polynomial regression, (A.1), to account for the biomodality as displayed in the histogram of Figure 3.In this way we can account for possible non-linearity, while still restricting the amount of variables. Since we have almost 3000 observations, the risk of overfitting is limited. After estimation, we also construct two-sided 95% confidence intervals for the OLS $\hat{y}$ (A.2).

2.3 Kernel density regression

Our non-parametric regression model can be written as (A.3), where the amount of daily traffic accidents is given by y, and x denotes the average temperature for the accidents on that day. We apply the Nadayara-Watson estimator (A.4) and limit ourselves to a Gaussian Kernel (A.5), since Cameron and Trivedi (2005) also state that bandwidth choice is more important than choice of kernel. To optimize the results, we try three different bandwidths: the plug-in Gaussian, Silverman and cross-validation (CV) bandwidth (A.6). The CV variant here is the bandwidth minimizing the RMSE, found using cross-validation. Assuming asymptotic normality, we can construct a confidence interval (A.7) for $\hat{m}(x)$ in all of these specifications. We estimate the bias by bootstrap.

3 Results

The estimated density functions using a Gaussian and Silverman bandwidth are plotted in Figure 3, along with Matlab's own density estimation function. Since for our dataset $min(s, iqr/1.349) = s$, the Silverman and Gaussian plug-in bandwidths are equivalent . We note that the non-parametric estimations match the original data quite well, with only Matlab's density function undersmoothing slightly.

Using this density function in a regression framework, the results of the kernel regressions are presented in Figure 4, along with their 95% confidence intervals. We see that the estimates of the non-parametric and parametric methods are very similar, with all of them finding a slight positive relationship between the temperature and the amount of daily collisions. Because the variance in our dataset is

so large and the relationship between our dependent and independent variable is weak, the confidence intervals for the OLS regression estimates are very wide. It seems that the non-parametric methods are better at dealing with this uncertainty, as we see that their confidence interval is much narrower.
For this dataset, the choice of bandwidth does not have a big impact on the results. However, since the confidence interval of the estimations using the cross-validation bandwidth is slightly narrower than that of the Gaussian/Silverman plug-in bandwidth, the former is marginally preferable in this case.

4 Conclusion

Concluding, we find that there is a small positive correlation between the temperature and the accident frequency. Parametric and non-parametric regression produce similar estimates for our data, but the confidence intervals of the kernel density regressions are more reliable, which makes a non-parametric approach preferable in this case. The bandwidth optimized using cross- validation leads to the best results, although the impact of changing the bandwidth is very small. Because of the large variance and weak correlation between the dependent and independent variables in our data, these findings can not be regarded as conclusive evidence and further research would be wanted. This research could for example look at the amount of daily precipitation instead of the temperature.

References

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Leard, B., Roth, K., et al. (2015). Weather, traffic accidents, and climate change. *Resources for the Future Discussion Paper*, 15–19.

Zou, Y., Zhang, Y., & Cheng, K. (2021). Exploring the impact of climate and extreme weather on fatal traffic accidents. *Sustainability*, *13*(1), 390.

A Appendix: Formulae

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 \tag{A.1}$$

$$CI_{95\%} = [X\hat{\beta} - 1.96s,\ X\hat{\beta} + 1.96s] \tag{A.2}$$

$$y_i = m(x_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon) \tag{A.3}$$

$$\hat{m}(x) = \frac{\frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x_i - x}{h}) y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x_i - x}{h})} \tag{A.4}$$

$$K(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2} z^2)) \tag{A.5}$$

$$\begin{aligned} \text{Plug-in: } h^* &= 1.3643 \cdot \delta \cdot N^{-1/5} \cdot N^{-1/5} s \\ \text{Silverman: } h^* &= 1.3643 \cdot \delta \cdot N^{-1/5} \cdot min(s, iqr/1.349) \\ \text{Cross-validation: } h^* &= min_h \ \frac{1}{N} \sum_{i=1}^{N} ((y_i - \hat{m}_{-i}(x_i))^2 \cdot \pi(x_i) \end{aligned} \tag{A.6}$$

$$CI_{95\%} : m(x_0) \in \left[\widehat{m}\,(x_0) - \widehat{\text{bias}}\,(x_0) \pm 1.96 \cdot \hat{\sigma}_\varepsilon \sqrt{\frac{1}{hN} \frac{\int K(z)^2 dz}{\hat{f}\,(x_0)}} \right] \tag{A.7}$$

B Appendix: Figures

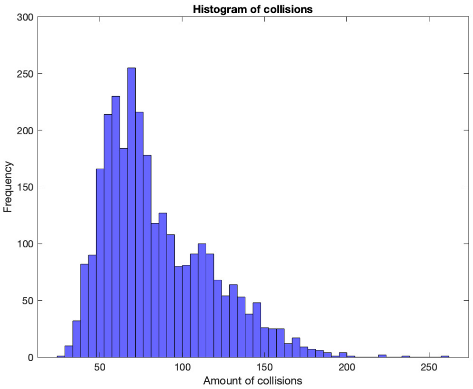

Figure 1: Histogram of the amount of daily collisions, using all data

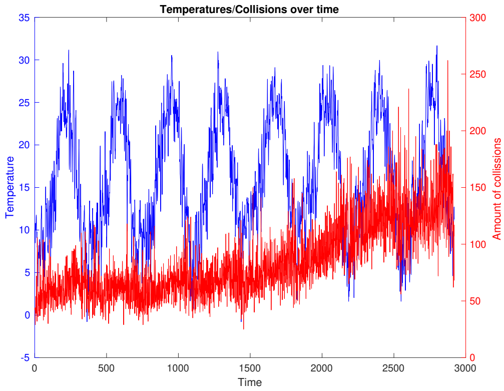

Figure 2: Development of daily average temperatures and daily number of collisions over time (in days), using all data

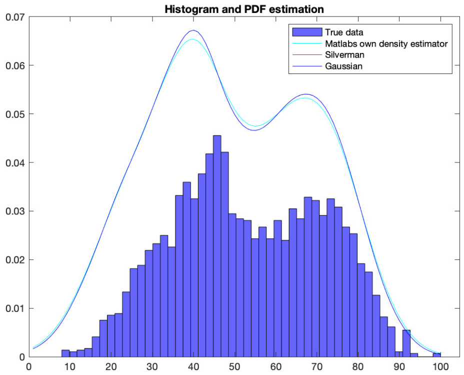

Figure 3: Kernel density estimation, compared to the histogram of actual data, conditional on no rain

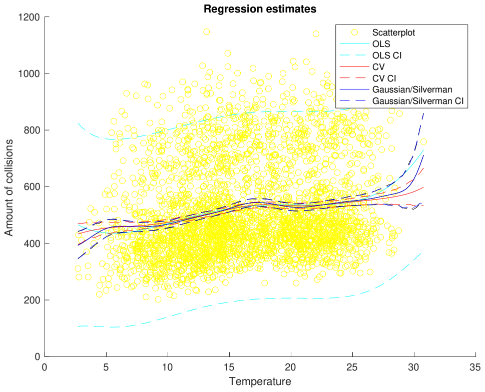

Figure 4: Regression estimates of parametric and non-parametric methods and their respective 95% two-sided confidence intervals, conditional on no rain