Faculty Economics & Business, Amsterdam School of Economics
University of Amsterdam

Density Estimation and Kernel Regression on US traffic data

Assignment AE2 Week 3: Semi- and Non-parametric methods

Vien Dinh, Derek Dirks & Jorrim Prins

11002115, 11029633 & 11038934

23-01-2020

Abstract

Non-parametric techniques offer estimation methods different from classical linear regression. In this research, the former is used to estimate the relation between traffic volume and temperature in Minnesota between 01-10-2016 and 30-09-2018. Characteristics of the techniques are analysed, along with a comparison to parametric methods. Traffic volume and temperature are concluded to be positively correlated.

Contents

1 Introduction

In this report, the relation between the temperature and hourly traffic volume on a highway in Minnesota is analysed using parametric and non-parametric methods. The data is measured on the westbound lane of the I-94 in a two year span strating October 1st 2016, consisting of 21195 observations of the temperature measured in Kelvin and traffic volume per hour. By applying parametric and non-parametric estimation techniques, the probability density of the traffic volume, the distribution of the traffic volume conditional on the temperature and optimal bandwidth methods are analysed.

This research aims to draw inference on the properties of the relation between traffic volume and temperature on the specified highway. The estimates following from the models in this paper may present useful information that could help reduce traffic jams or determine optimal maintenance periods for the highway.

Sathiaraj, Punkasem, Wang, and Seedah (2018) found that a lower temperature leads to as much as a 10% drop in traffic volume for some hours of the day in Atlanta, Georgia. In Scotland, a higher than expected temperature corresponds to an increase in traffic volume and a lower than expected temperature corresponds to a decrease in traffic volume (Al Hassan & Barker, 1999). Therefore, expect traffic volume to be positively correlated with the temperature.

2 Models & Techniques

2.1 Density estimation

To obtain a better understanding of the relation between automobile traffic and the weather, the non-parametric densities are estimated by various techniques. Solely the Gaussian kernel is applied to evaluate these non-parametric models, as this is the simplest implementation. Furthermore, Cameron and Trivedi (2005) state that "bandwidth choice is much more important than kernel choice". Density estimations for the plug-in, Silverman and optimal bandwidths are compared and respectively look as follows:

$$h^* = a\delta N^{-0.2}s, \quad a\delta N^{-0.2}min(s, iqr/1.349), \quad \delta\left(\int\left(f''\left(x_0\right)^2 dx_o\right)^{-0.2} N^{-0.2}\right) \tag{1}$$

with $a = 1.3643$ and $\delta = 0.7764$ for Gaussian kernels.

These density estimates are compared with maximum likelihood estimates of a similar distribution, which is a bimodal one in this case.

2.2 Non-parametric regression

For the kernel regression the following model is estimated:

$$y_i = m(x_i) + \epsilon_i, \quad \epsilon_i \sim iid[0, \sigma^2] \tag{2}$$

In this model, y_i is the hourly westbound traffic volume and x_i is the explanatory variable, which exists of the temperature in Kelvin. The relation between these variables can be estimated by the kernel regression estimator shown in (3), where again Gaussian kernels are used. Parallel to this regression the heteroskedastic-robust 95% two-sided confidence interval is calculated with the use of (4).

$$\hat{m}\left(x_0\right) = \frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)} = \sum_{i=1}^{N} w_{i0,h} y_i \tag{3}$$

$$CI_{95\%} = \hat{m}\left(x_0\right) \pm 1.96\hat{s}_0^2 = \hat{m}\left(x_0\right) \pm 1.96 \sum_i w_{i0,h}^2 \hat{\varepsilon_i^2} \tag{4}$$

The estimates of the kernel regression are compared to a parametric OLS model and corresponding confidence intervals, with simple but similar polynomial features. Accordingly, the cubic model in (5) is used.

$$y_i = \alpha_i + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i \tag{5}$$

2.3 Bandwidths

Whereas kernel density estimations usually provide reliable estimations with plug-in bandwidths, kernel regression might be improvable by, for example, cross-validation. This paper tries to find the optimal bandwidth for the kernel regression by using generalised cross-validation which optimises (6), in which the final term is a penalty function and $\pi(x_i)$ is defined in two different ways. The outcomes are subsequently optimised by "eyeballing" the curve.

$$PV(h) = \sum_{i=1}^{N} \left(y_i - \hat{m}\left(x_i\right)\right)^2 \pi\left(x_i\right) \left(1 - w_{ii,h}\right)^2 \tag{6}$$

3 Results

3.1 Density estimation

In figure 1, the kernel density function using a Gaussian kernel with the three previously mentioned techniques and a bimodal MLE is plotted. Since $min(s, iqr/1.349) = s$ for this dataset, the plug-

in and Silverman bandwidths are equal. Furthermore, the figure shows that the right peak is correctly estimated using ML, while the left peak is oversmoothed as it attempts to fit the left and middle peak in one. This suggests that a bimodel maximum likelihood model is not the correct specification for this data.

3.2 Regression analysis

In figure 2, the results of kernel and OLS regressions of traffic volume per hour on the temperature are plotted with their respective 95% confidence intervals. For temperatures between 260 and 290 degrees, kernel regression shows that traffic volume increases marginally with the temperature. For extreme temperatures the slope of the relation becomes steeper, downwards for extreme low temperatures and upwards for extreme high temperatures. The OLS regression show similar results, but is obviously disregarding a lot of small changes in the curve. End-point bias is also apparent for both curves as confidence intervals widen towards the end-points. This is caused by the low number of observations in the tails of the data.

3.3 Bandwidth analysis

Figure 3 shows the generalised cross validation results, where the PV-value is plotted as a function of the bandwidth for a Gaussian kernel. The first plot uses $\pi(x_i) = \frac{1}{N}$ (no end-point bias correction) and the second plot uses $\pi(x_i) = \frac{1}{0.8N}$ (and only the middle 80% of the data), both show globally increasing functions and optimal values for h close to zero. Figure 4 shows how this value for the bandwidth would estimate the kernel regression.

4 Conclusion

To conclude, non-parametric regression offers methods without assumptions to estimate densities and compute estimators that are, in this case, closer to the actual data than parametric counterparts. Kernel selection is not applied in this research, as bandwidth selection is proven to be more important. Various bandwidth values are exercised, but optimisation methods seem to result in undersmoothing estimations. Regular plug-in bandwidths are used for densities and estimators and show a positive correlation between hourly traffic volume and temperature, with increasing slopes for extreme values of the temperature. This is in accordance with earlier mentioned research.

5 Appendix

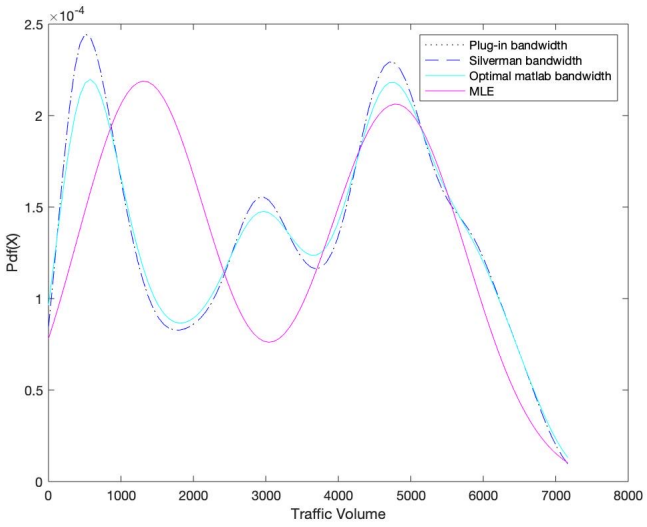

Figure 1: Probability Density Function (PDF) estimation of the traffic volume

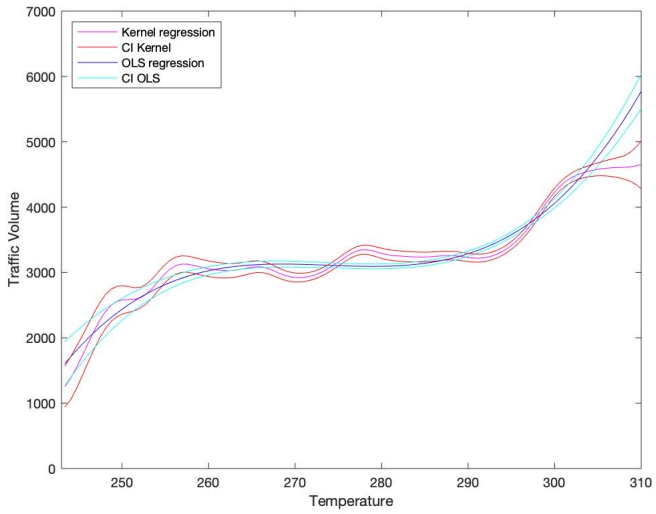

Figure 2: Kernel and OLS regression of traffic volume conditional on temperature with the corresponding 95% confidence interval

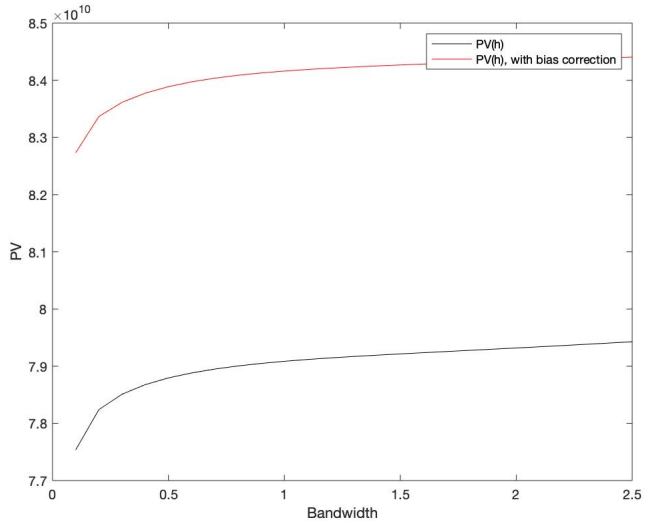

Figure 3: Generalised Cross Validation measure (PV) for different values of the bandwidth (h)

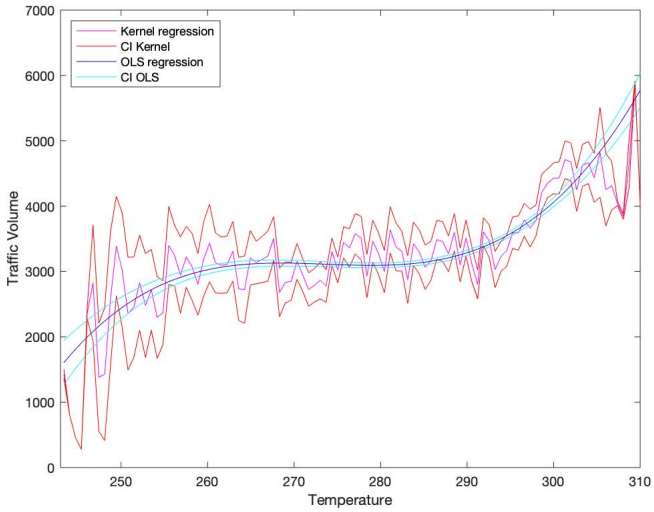

Figure 4: Kernel and OLS regression of traffic volume conditional on temperature with the corresponding 95% confidence interval, with bandwidth close to 0

References

Al Hassan, Y., & Barker, D. J. (1999). The impact of unseasonable or extreme weather on traffic activity within lothian region, scotland. *Journal of Transport Geography*, *7*(3), 209–213.

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

Sathiaraj, D., Punkasem, T.-o., Wang, F., & Seedah, D. P. (2018). Data-driven analysis on the effects of extreme weather elements on traffic volume in atlanta, ga, usa. *Computers, Environment and Urban Systems*, *72*, 212–220.