Non-Parametric Estimation of Geographical Relative Risk Functions

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at Massey University, Palmerston North

New Zealand

W. T. P. Sarojinie Fernando

December 2012

Abstract

The geographical relative risk function is a useful tool for investigating the spatial distribution of disease based on case and control data. The most common way of estimating this function is using the ratio of spatial kernel density estimates constructed from the locations of cases and controls respectively. This technique is known as the density ratio method. The performance of kernel density estimators depends on the choice of kernel and the smoothing parameter (bandwidth). The choice of kernel is not critical to the statistical performance of the method but the bandwidth is crucial. Different bandwidth selectors such as least squares cross validation (LSCV) and likelihood cross validation (LCV) are chosen to control the degree of smoothing during the computation of the density ratio estimator.

An alternative way of estimating this relative risk function is local linear regression approach. This deserves consideration since the density ratio estimator can be less natural when the relative risk has a global trend, as one might expect to see when there is a line source of risk such as a polluted river or a road. The use of local linear regression for estimation of log relative risk functions *per se* has not been examined in any detail in the literature, so our work on this methodology is a novel contribution. A detailed account of local linear approach in the estimation of log relative risk function is provided, consisting of an analysis of asymptotic properties and a method for computing tolerance contours to emphasize the regions of significantly high risk. Data driven bandwidth selectors for the local linear method including a novel plug-in methodology is examined.

A simulation study to compare the performance of density ratio and local linear estimators using a range of data-driven bandwidth selectors is presented. The analysis of two specific data sets is examined.

The estimation of the spatial relative risk function is extended to spatio-temporal estimation through the use of suitable temporal kernel functions, since time-scale is an important consideration when estimating disease risk. The extended version of the kernel density estimation is applied here to compute the unknown densities of the spatio-temporal relative risk function. Next we investigate the time derivatives of the space-time relative risk function to see how the disease change with time. This discussion provides novel contributions with the introduction to time derivatives of the relative risk function as well as asymptotic methods for the computation of tolerance contours to highlight subregions of significantly elevated risk. LSCV and subjective bandwidths are used to compute these estimators since it performs well in density ratio method. The analysis on a real application to foot and mouth disease (FMD) of 1967 outbreak is employed to illustrate these estimators.

The relative risk function is investigated when the data include a spatially varying covariate. The discussion produces the introduction to generalized relative risk function in two ways and also asymptotic properties of estimators for both cases as novel works. Generalized kernel density estimation is used to replace the unknown densities in the relative risk function. Asymptotic theories are used to compute tolerance contours to identify the areas which show high risk. LSCV bandwidth selector is described in this estimation process providing the implicit formulae. We illustrate this methodology on data from the 2001 outbreak of FMD in the UK, examining the effect of farm size as a covariate.

Acknowledgements

I am sincerely and heartily grateful to my supervisor, Martin Hazelton, for his continuous support and guidance throughout this research. I am sure this would have not been accomplished without his help. I would also like to thank Martin for being patient specially when I was making less progress. Finally I would like to express my heartfelt gratitude to Martin for his valuable comments and suggestions to direct this dissertation a successful one.

I am thankful to Ganes Ganesalingam for his continuous guidance in numerous ways as being my co-supervisor in the early years of my research. My thanks go to Jonathan Marshall for being my co-supervisor in my final year due to the replacement of Ganes.

I would like to acknowledge all the members in the Statistics group for their valuable comments, and also colleagues in the Institute office of Fundamental Sciences. My special thanks go to IFS for providing me a scholarship at the end of Graduate assistant position and also to Steve Haslett for providing me funds to pay tuition fee afterwards. I should thank my postgraduate colleagues at the Statistics group, Massey University for their encouragement and providing me a pleasing environment.

My special thanks go to two important people. Sarath Kulatunga, a Senior Professor at the University of Kelaniya in Sri Lanka, helped me to acquire a thorough statistical knowledge while I was working towards my M.Phil. with him. I would also like to express my gratitude to Priyantha Wijayatunga, a lecturer at Umea University

in Sweden, who encouraged me to apply to this Graduate assistant Ph.D. position. I take this opportunity to thank to all staff members at the Department of Mathematics, University of Kelaniya. My thanks also go to staff members at the faculty of Applied sciences, specially to my colleagues at the Department of Mathematics, Wayamba University of Sri Lanka for making a friendly environment while I was working there as a lecturer.

Of course, I am indebted to my lovely husband Kithsiri Fernando for his love, support, motivation and constant patience which have taught me so much about sacrifice during this journey. Without him this work would never have come into existence. I am also grateful to my son, Kaveesha Fernando and my daughter, Gihara Fernando for their profound understanding during this period. I missed them so much. I am extremely sorry for the time we spent apart.

I would like to acknowledge my lovely mum, Janet for loving, encouraging me always, believing in me, in all my efforts. I also express my gratitude to my late dad, Julian for making a dream in my mind to be a successful academic when I was a child. I also like to acknowledge my brother and sister for their continuous support. My special remind goes to my mother-in-law, Mary who unfortunately passed away while I was reading for my Ph.D.

Last but not least, I would like to thank my ever loving God for helping and guiding me to be successful throughout my life.

Table of Contents

List of Figures

List of Tables

Notation

$\mathbf{I}$	2×2 identity matrix.
f	Case density function
g	Control density function
r	Relative risk function
ρ	Log relative risk function
$supp(f)$	Support of f
p	Degree of the polynomial, conditional probability and size of covariate.
h	Fixed bandwidth
h_1	Fixed bandwidth for case data.
h_2	Fixed bandwidth for control data.
h_{OS}	Over smoothing bandwidth.
q	Edge correction factor.
q_1	Edge correction factor based on case data.
q_2	Edge correction factor based on control data.
$\mathbf{H}$	bandwidth matrix
Σ	Covariance matrix.
λ	temporal bandwidth
K	Unscaled kernel function
K_h	Scaled kernel function, scaled with h.
$\mu_2(K)$	Second central moment of K
$R(f)$	$\int_{\mathcal{R}} f(\mathbf{x})^2 d\mathbf{x}$
$R(K)$	$\int_{\mathcal{R}} K(\mathbf{z})^2 d\mathbf{z}$
$R(f'')$	$\int_{\mathcal{R}} f''(\mathbf{z})^2 d\mathbf{z}$
f''	Second derivative of the density f.
L	Univariate probability density function.
$\mathcal{D}_{\{}(\mathbf{x})$	First derivative of f with respect to $\mathbf{x}$.
$\mathcal{H}_{\{}(\mathbf{x})$	Hessian of f with respect to $\mathbf{x}$.
$\phi_\sigma(\mathbf{x} - \mu)$	Multivariate normal density.
μ	Mean.

σ^2	Variance.
$\mathcal{R}$	Geographical region
$\|\mathcal{R}\|$	Area of the geographical region $\mathcal{R}$.
∇^2	Laplacian operator.
DR	Density Ratio
LL	Local Linear
RR	Relative risk
ISE	Integrated Squared Error
MSE	Mean Squared Error
AMSE	Asymptotic Mean Squared Error
MISE	Mean Integrated Squared Error
WMISE	Weighted Mean Integrated Squared Error
MIAE	Mean Integrated Absolute Error
AMISE	Asymptotic Mean Integrated Squared Error
PI	Plug-In bandwidth selector
GAM	Generalized additive model.
LSCV	Least Squares Cross Validation
LCV	Likelihood Cross Validation
SE	Standard Error
x_i	scalar.
$\mathbf{x}_i$	A vector of dimension 2
$x_1, ..., x_n$	Random scalar sample of size n.
$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$	a random sample of size n
t	Time of occurence
T	Time interval
$\|T\|$	Length of time interval
FMD	Foot and mouth disease
KDE	Kernel density estimate
$\hat{f}(x;h)$	Fixed kernel density estimate constructed from univariate data x.
$\hat{f}(\mathbf{x};h)$	Fixed kernel density estimate constructed from bivariate case data $\mathbf{x}$.
$\hat{g}(\mathbf{x};h)$	Fixed kernel density estimate constructed from bivariate control data $\mathbf{x}$.
$\hat{r}(\mathbf{x};h)$	Fixed bandwidth relative risk estimate.
$\hat{\rho}(\mathbf{x};h)$	Fixed bandwidth log relative risk estimate.
$\hat{\rho}_{LL}$	LL estimator of log relative risk function.
$\hat{\rho}_{LC}$	Local constant estimator.
$\hat{f}^{-i}(\mathbf{x};h)$	Leave-one-out estimate based on case data
$\hat{g}^{-i}(\mathbf{x};h)$	Leave-one-out estimate based on control data
h_{opt}	The optimal bandwidth
h_{MISE}	MISE- optimal bandwidth
$\hat{h}_{PI}$	Plug-in bandwidth selector
$\hat{h}_{LSCV}$	LSCV bandwidth selector
$\hat{h}_{LCV}$	LCV bandwidth selector
$\hat{h}_{LL.LCV}$	LCV bandwidth for LL estimator.
$\hat{h}_{LL.PI}$	PI bandwidth for LL estimator.

$\hat{h}_{DR.CV}$	LSCV bandwidth for DR estimator.
$\hat{g}_p$	Pooled estimator.
n_1	Case sample size
n_2	Control sample size
n	Equal to $n_1 + n_2$
π	Equal to $\frac{n_1}{n_1+n_2}$
$\bar{L}$	The likelihood function
$\bar{L}'$	The first derivative with respect to β
$\bar{L}''$	The second derivative with respect to β
P, Q	Polynomials.
$p(\mathbf{x})$	Conditional probability for a given point $\mathbf{x}$.
y_i	Binary variable.
Cov	Covariance.
tr	Trace.
δ	A small positive number.
δ_0	A small positive number.
ϵ	A small constant.
$\bar{f}$	Edge corrected kernel density estimate constructed from case data.
$\bar{g}$	Edge corrected kernel density estimate constructed from control data.
H_0,H_0^1,H_0^2	Null hypotheses.
H_1,H_1^1,H_1^2	Alternative hypotheses.
$\mathbb{E}$	Expectation.
θ	A parameter.
Z	Test statistic.
$W_{\mathbf{x}}$	$n \times n$ diagonal matrix at $\mathbf{x}$.