CS-BIGS 2(2): 120-126 http://www.bentley.edu/csbigs/vol2-2/willie.pdf


# Estimating Determinants of Healthcare Utilisation: A Nonparametric Approach

**Mncedisi Michael Willie**
*Council for Medical Schemes, South Africa*

*The goal of this paper is to propose a nonparametric analysis applied to medical schemes data. Because medical schemes data are aggregated and have heavy tailed densities, we consider nonparametric density estimation, which overcomes limitations related to aggregation bias and heavy tails in the data. We demonstrate the usefulness of this class of models by using an empirical example that seeks to predict healthcare use. This case study can be used by researchers in health-care as well as in intermediate course in data analysis in order to highlight the importance of distribution-free models in the context of a real data set.*

## 1. Introduction

Medical schemes are the dominant vehicle for providing health care insurance in the private sector in South Africa. Medical schemes reimburse their members for actual expenditure on health. They operate on a not-for-profit basis and are essentially mutual societies, governed under the Medical Schemes Act (Act 131 of 1998) (McLeod and Ramjee 2007). The Medical Schemes Act defines the 'business of a medical scheme' as the business of undertaking liability in return for a contribution in order to make provision for obtaining any 'relevant health service'. Schemes may choose to be restricted membership schemes if attached to a large employer, union or other defined group, but all others are open schemes that must freely admit anyone who applies (Republic of South Africa 1998).

Medical schemes data report on healthcare services provided to members as well as reimbursement for such services. The type of healthcare delivered to a specific individual depends on her/his own health characteristics. Because of characteristics of schemes operating in the South African health care system as well as differences based on individual healthcare needs, parametric models are sometimes too restrictive for medical schemes data which tend to feature heavy tails (Maurer 2002).

This article recommends nonparametric models as a tool for modelling the interplay of health characteristics in determining health-care utilisation. In application to medical schemes data, it is important to use a method robust to heavy-tailed densities. Nonparametric density estimation overcomes the trend surface limitations of aggregation bias by using data points directly (Silverman 1986).

We demonstrate with the help of an empirical example the insights that nonparametric methods may deliver in estimating the model for utilisation of GPs (General Practitioners) services. Bivariate nonparametric density estimation has been reported as a powerful tool for revealing complicated relationships between two variables (Wilkinson 1994) so we employ this method to perform data exploration on healthcare utilisation. We also demonstrate an application of quantile regression in order to see the interplay of health characteristics in determining health-care utilisation. The results of both these approaches are presented and results discussed.

This article is developed in the following way. In section 2 we give some background on the data used in this paper and on the collection of medical schemes data and provide definitions of the variables in the dataset. Sections 3 and 4 explain our methodology. We then present the results and findings in section 5, and finally give some concluding remarks.

## 2. Data and Methods

The study included open schemes (where a scheme freely admits anyone who apply) and restricted schemes (where a member is attached to a large employer, union or other defined group) registered with the Council for Medical Schemes (CMS). The Council for Medical Schemes is an autonomous statutory body, established by the Parliament of South Africa to provide supervision over medical schemes (Council for Medical Schemes 2006a).

Data for the analysis was sourced from the Annual Statutory Return (ASR) for 2007, which contains information on the demographic profile of beneficiaries' utilisation of services and on expenditure on healthcare services (Council for Medical Schemes 2006b). A random sample of 56 (28 open and 28 restricted) schemes was selected for the analysis on the basis of the completeness of the data available for the schemes. The data were electronically submitted to the Council via a secure internet portal. Schemes final data submissions were downloaded onto Microsoft excel spreadsheets. The data were then imported to both Stata and SAS 9.1 for analysis.

The objective of this article is to suggest nonparametric models as a tool for modelling the interplay of health determinants in determining health care utilisation. The goal of the analysis was to investigate the association of healthcare utilisation with healthcare characteristics (predictors of healthcare). This was achieved by employing two nonparametric methods: bivariate nonparametric density estimation and quintile regression. We now define variables considered in the analysis.

## 3. Kernel Density Estimation (KDE)

Kernel density estimation is probably the most widely used method for smoothing one-dimensional or multidimensional sample data into a continuous probability density function. Rosenblatt (1956) first explicitly introduced the kernel estimate (Rosenblatt 1956), defined in d dimensions as:

**Table 1.** List of variables used

| Variable | Description |
|---|---|
| **Outcome/ response variable** | |
| *gpv* | Total number of visits to a general practitioner per year per beneficiary |
| **Covariates** | |
| *scheme type* | open, and restricted schemes |
| *scheme size* | Small, medium, and large schemes |
| *Female_visists* | Number of visits to a general practitioner per female beneficiary |
| *Male_visists* | Number of visits to a general practitioner per male beneficiary |
| *Hypertension* | Number of reported cases of hypertension ( per 1000 beneficiaries) |
| *Hyperlipidaemia* | Number of reported cases of Hyperlipidaemia ( per 1000 beneficiaries) |
| *Asthma* | Number of reported cases of Asthma ( per 1000 beneficiaries) |
| *Coronary Artery Disease* | Number of reported cases of Coronary Artery Disease ( per 1000 beneficiaries) |
| *HIV* | Number of reported cases of HIV ( per 1000 beneficiaries) |
| *Hypothyroidism* | Number of reported cases of Hypothyroidism ( per 1000 beneficiaries) |
| *Epilepsy* | Number of reported cases of Epilepsy ( per 1000 beneficiaries) |
| *Diabetes Mellitus Type 1* | Number of reported cases of Diabetes Mellitus Type 1 ( per 1000 beneficiaries) |
| *Diabetes Mellitus Type 2* | Number of reported cases of Diabetes Mellitus Type 2 ( per 1000 beneficiaries) |
| *Cardiac Failure* | Number of reported cases of Cardiac Failure ( per 1000 beneficiaries) |

Note: The healthcare characteristics variables listed above were selected merely for illustrative purposes.

$$\hat{f}(x) = \left(nh^d\right)^{-1} \sum_{i=1}^{n} K\left((x - x_i)/h\right) \quad (1)$$

where K is the kernel function, h is the smoothing factor or window width, n is the number of data points, and $x_i$ are the data points.

The bivariate kernel density estimator is a special case of (1) defined by

$$\hat{f}(x) = \left(nh^2\right)^{-1} \sum_{i=1}^{n} K\left((x - x_i)/h\right). \quad (2)$$

Note that is it usually assumed that $\int K(x)dx = 1$; typically $K(x)$ will be a radially symmetric unimodal probability density function, for example, the standard bivariate normal density function:

$$K(x)=\frac{1}{2\pi}\exp\left(-\frac{1}{2}x'x\right)\cdot \quad (3)$$

## 4. Quantile Regression

Quantile regression (see for example Chen 2005 and Koenker and Hallock 2001) generalizes the concept of a univariate quantile to a conditional quantile given one or more covariates. For a random variable Y with cumulative distribution function

$$F(y)=\text{Prob}(Y\le y), \quad (4)$$

the $\tau^{th}$ quantile of Y is defined as the inverse function

$$Q(\tau)=\inf\{y:F(y)\ge\tau\} \quad (5)$$

where $0<\tau<1$. In particular, the median of Y is $Q(1/2)$. Given a random sample $\{y_1,\ldots,y_n\}$ from Y, it is well known that the sample median is the value of $\xi$ which minimizes of the sum of absolute deviations

$$\sum_{i=1}^{n}|y_i-\xi| \quad (6)$$

Likewise, the general $\tau^{th}$ sample quantile $\xi(\tau)$, which is the analogue of $Q(\tau)$, may be formulated as the solution of the optimization problem:

Find $\xi$ to minimize

$$\min_{\xi\in R}\sum_{i=1}^{n}\rho_\tau(y_i-\xi) \quad (7)$$

where $\rho_\tau(z)=z(\tau-I(z<0))$ for $0<\tau<1$. Here $I(.)$ denotes the indicator function; note that $\rho_{1/2}(z)=|z|/2$.

Just as a sample mean, which minimises the sum of squared residuals

$$\hat{\mu}=\arg\min_{\beta\in R^p}\sum_{i=1}^{n}(y_i-\mu)^2 \quad (8)$$

can be extended to the linear conditional mean function $E(Y|X=x)=x'\beta$ yielding the least-squares estimator

$$\hat{\beta}=\arg\min_{\beta\in R^p}\sum_{i=1}^{n}(y_i-x_i'\beta)^2, \quad (9)$$

the linear conditional quantile function, $Q(\tau\mid X=x)=x'_i\,\beta(\tau)$ can be estimated by solving

$$\hat{\beta}(\tau)=\arg\min_{\beta\in R^p}\sum_{i=1}^{n}\rho_\tau(y_i-x_i'\beta), \quad (10)$$

for any quantile $\tau\in(0,1)$. The quantity $\beta(\tau)$ is called the $\tau^{th}$ regression quantile estimator.

## 5. Findings and Discussion

This section presents an empirical example to bring the above discussed concepts to life, and also demosntrates the applicability of such non-parametric models to the data on utilisation of general practitioners. The results are based on the analysis of a random sample of 56 registered schemes with complete utilisation data. Twenty-eight of the schemes were open schemes and the remaining twenty-eight were restricted schemes; this sample represented 45 percent of the 124 registered schemes in 2006. The number of beneficiaries from the 56 schemes represented was 6.2 million beneficiaries, which represented 87 percent of all beneficiaries.

We note that there was no significant difference between the age profile of beneficiaries in open and restricted schemes (the p-value for the equality of mean ages was 0.7932). The average age of beneficiaries in open and restricted schemes was 28 and 29 years respectively.

Tables 2 and 3 display basic descriptive statistics for some of the variables used in the analysis, for open and restricted schemes respectively. There was no significant difference in the mean number of visits to a general practitioner for open and restricted schemes (p-value=0.83). The average number of doctor visits in 2006 was 3.0 and 3.1 visits per beneficiary per year for open and restricted schemes respectively.

An additional noteworthy feature of the data is the presence of large differences between open and restricted schemes in terms of prevalence rates. Selected conditions such as Hypertension (p=0.016), Asthma (p=0.039), Epilepsy (p-value=0.02) and Diabetes Mellitus Type 2 (p=0.015) were significantly higher in restricted schemes compared to open schemes.

### 5.1 illustrations using the kernel density estimation

The most common two-dimensional continuous variable data display is the scatter plot. Figure 1 shows an example of such a scatter plot, using number of visits to a general practitioner for female and male beneficiaries on the y and x axes respectively. We note in passing that there were no significant difference between the mean number of visits to a general practitioner for male and female beneficiaries.

The purpose here is to show how one can smooth data points displayed in Figure 1 in order to reveal hidden structures in the data. Figure 1 was generated with the statistical software package Stata. Figures 2 -4 were produced with SAS 9.1. SAS code for producing the graphs in Figures 2-4 is given in the appendix.

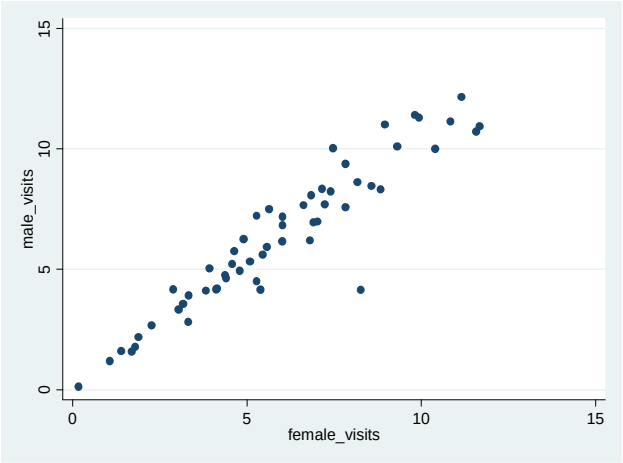


**Figure 1.** Scatter plot of male and female visits to GP per beneficiary

Figure 2 shows the same scatter plot displayed in Figure1, but in the form of a contour plot of a two dimensional kernel. In Figure 2 we employed a rule-of-thumb bandwidth as a smoothing parameter; however, we note that this tends to over-smooth the data. The figure reveals that most of the data points are concentrated around 5 visits for both males and females. But if we enhance this figure and adjust the bandwidth we get a much more informative picture.

Figure 3 shows how powerful this smoothing and data display can be after adjusting the bandwidth to 0.5 for both variables. The figure also displays some evidence of where most of the data points are concentrated, but we now note that there are several clusters in the data which were not visible in Figure 2.

**Table 2.** Baseline characteristics for the open schemes

| Variable | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|
| average age per beneficiary | 32.35 | 4.31 | 26.39 | 45.74 |
| Total number of visits to a GP per beneficiary | 2.96 | 1.2 | 0.73 | 5.28 |
| Log (Total number of visits to a GP) | 12.26 | 1.35 | 9.72 | 15.39 |
| **Number of episodes ( per 1000 beneficiaries)** | | | | |
| Hypertension | 65.59 | 59.05 | 3.6 | 215.9 |
| Hyperlipidaemia | 36.96 | 50.18 | 1.4 | 255.6 |
| Asthma | 19.72 | 13.29 | 1.1 | 50.6 |
| Coronary Artery Disease | 15.79 | 23.78 | 0 | 106 |
| HIV | 10.09 | 20.26 | 0 | 87 |
| Hypothyroidism | 11.14 | 13.35 | 0 | 56.4 |
| Epilepsy | 5.17 | 3.53 | 0.2 | 14.8 |
| Diabetes Mellitus Type 1 | 4.5 | 3.23 | 0.2 | 13.4 |
| Diabetes Mellitus Type 2 | 16.42 | 11.65 | 0.4 | 36.2 |
| Cardiac Failure | 4.13 | 4.85 | 0 | 16.7 |

**Table 3:** Baseline characteristics for the restricted schemes

| Variable | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|
| Average age per beneficiary | 32.49 | 4.59 | 26.19 | 44.64 |
| Total number of visits to a GP per beneficiary | 3.14 | 1.66 | 0.11 | 5.98 |
| Log( Total number of visits to a GP) | 11.3 | 1.36 | 7.52 | 14.68 |
| **Number of episodes ( per 1000 beneficiaries)** | | | | |
| Hypertension | 104.96 | 75.05 | 2.3 | 300.8 |
| Hyperlipidaemia | 45.4 | 31.98 | 0 | 126.8 |
| Asthma | 26.66 | 15.58 | 0.9 | 75.3 |
| Coronary Artery Disease | 16.97 | 16.65 | 0 | 86.2 |
| HIV | 4.65 | 5.32 | 0 | 21.3 |
| Hypothyroidism | 16.58 | 13.02 | 0 | 39.9 |
| Epilepsy | 7.23 | 4.06 | 0.2 | 19.4 |
| Diabetes Mellitus Type 1 | 6.05 | 4.89 | 0.1 | 24.4 |
| Diabetes Mellitus Type 2 | 24.4 | 15.02 | 0.5 | 57.2 |
| Cardiac Failure | 6.12 | 8.15 | 0 | 34 |

The density estimates can also be presented in the form of perspective plots as shown in Figure 4.

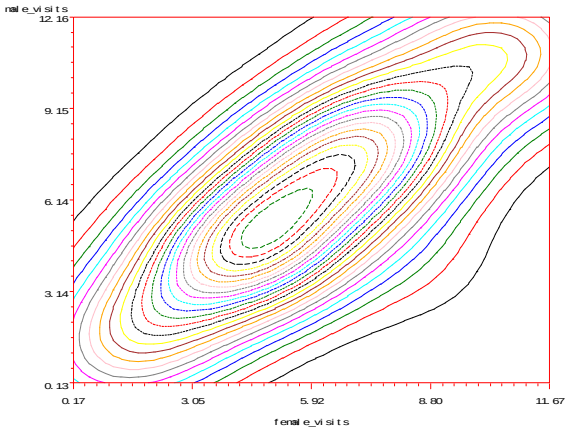


**Figure 2.** Bivariate kernel density contour plot of male and female visits to a GP per beneficiary with default bandwidths

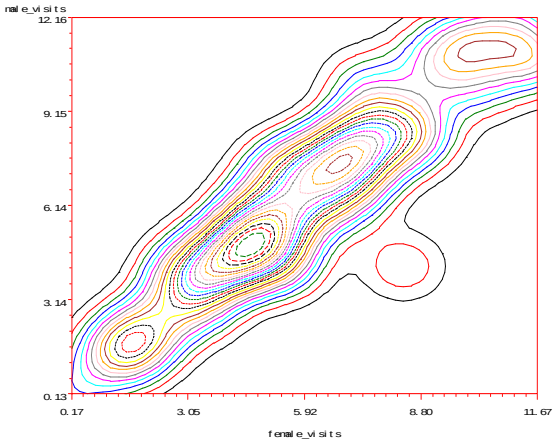


**Figure 3.** Bivariate kernel density contour plot of male and female visits to a GP per average beneficiary with bandwidths $h_x = 0.5$ and $h_y = 0.5$.

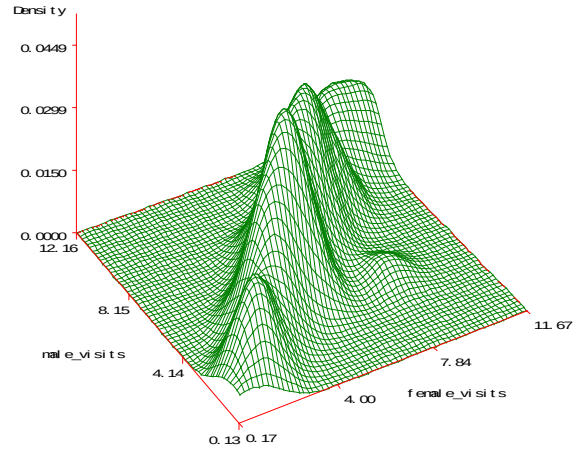


**Figure 4.** Three dimensional perspective plot of bivariate kernel density for male and female visits to a GP, with bandwidths $h_x = 0.5$ and $h_y = 0.5$.

The probability contours of the bivariate kernel density estimates of male and female visits to general practitioners revealed hidden structures in the data. The presence of clear clusters in the data illustrate that an individual healthcare characteristic may determine healthcare needs. In the next section we explore the use of a quantile regression model to test for associations between healthcare characteristics and healthcare use.

## 5.2 Illustration using quantile regression estimation

The analysis given in this section estimates how healthcare use depends on healthcare characteristics. A stepwise quantile regression approach was employed to add and subtract predictors from the model using a significance level of 5 percent for the threshold to either add or subtract predictors from the model. This resulting model was then used as our baseline model. The quantile regression procedure yielded the estimates presented in table 4.

**Table 4.** Stepwise quantile regression estimates (the median quantile was used)

| gpv | Coef. | Std. Err. | t | P>t | [95% Conf.] | |
|---|---|---|---|---|---|---|
| Hypertension | 0.014 | 0.006 | 2.510 | 0.015 | 0.003 | 0.025 |
| Cardiac Failure | 0.077 | 0.038 | 2.050 | 0.046 | 0.002 | 0.153 |
| Asthma | 0.089 | 0.025 | 3.570 | 0.001 | 0.039 | 0.139 |
| Epilepsy | -0.378 | 0.126 | -3.000 | 0.004 | -0.632 | -0.125 |
| Hypothyroidism | -0.087 | 0.024 | -3.570 | 0.001 | -0.136 | -0.038 |
| Constant | 2.953 | 0.410 | 7.210 | 0.000 | 2.130 | 3.776 |

The estimates of the model can be interpreted as follows. We note that the estimated median number of visits to a general practitioner can be obtained from the equation estimated by the model. The median number of visits to a general practitioner per year is 2.95 for schemes with no reported cases of the health care characteristics hypertension, cardiac failure, asthma, epilepsy or hypothyroidism (the constant in the model). Each additional case of hypertension per 1,000 beneficiaries implies an increase of 0.014 in the estimated median number of visits, when the prevalences of other characteristics are held constant (ceteris paribus); similarly each additional cardiac failure and asthma case per 1,000 beneficiaries implies an increase of 0.077 and 0.089 respectively in the estimated median number of visits, ceteris paribus. Each additional epilepsy and hypothyroidism case per 1,000 beneficiaries implies a decrease of 0.378 and 0.087 respectively in the estimated

median number of visits to a general practitioner, ceteris paribus.

These results affirm that there is a significant relationship between health care characteristics and health care need. However the results presented here should be interpreted with caution due to the relatively small sample size of the data set, but also to the fact that the prevalences may very well be correlated for the five different characteristics, so that it may be difficult for the ceteris paribus condition to hold.

## 6. Concluding remarks

The two methods presented in this article demonstrated how one can use the non-parametric approach to model healthcare data, particularly medical schemes data. We have shown that bivariate densities can be useful in detecting patterns in the data. We further explored a selected group of healthcare characteristics and assessed their association to healthcare utilisation. Our quantile (median) regression model revealed that there is a significant relationship between health care utilisation and healthcare characteristics such as hypertension, cardiac failure, asthma, epilepsy, and hypothyroidism.

The kernel density estimation and quintile regression model presented here provide attractive methods for modelling health determinants of healthcare use; however we must note that it would be most useful to use a more complete set of predictors, applied to a larger sample of data.

## Acknowledgements

I would like to thank Patrick Matshidze and Aleksandra Serwa for their continued support and encouragement. I am also grateful to all the units at the Council for Medical Schemes.

## Appendix: Code used to produce the graphs.

Figure 1 was generated in Stata 9 and Figures 2-4 were generated in SAS, the code is as follows:

```
proc kde data=gp  out=kdeout;
var male_visits female_visits;
run;

proc gcontour data=kdeout;
plot male_visits*female_visits=density/
nlevels=25 nolegend /*annotate=anno*/;
*title' Figure 2: Scatter plot of male and female
visits to GP per beneficiary with kernel smooth';
run;

proc kde data=gp  out=kdeout2 bwm=0.5,0.5;
var male_visits female_visits ;

run;
```

```
proc gcontour data=kdeout2 ;
plot male_visits*female_visits=density/
nlevels=25 nolegend ;
*title 'Figure 3: Scatter plot of male and female
visits to GP per average beneficiary with kernel
smooth';
run;

proc g3d data=kdeout2;
plot male_visits*female_visits=density/rotate=30
tilt=45;
*title'Figure 4: Perspective plot of estimated
bivariate density for male and female visits to a GP';
run;
```

## REFERENCES

Chen, C. (2005). An Introduction to Quantile Regression and the QUANTREG Procedure. Paper 213-30, SUGI30 Proceedings, available at http://www2.sas.com/proceedings/sugi30/213-30.pdf

Council for Medical Schemes (2006a). Pretoria, South Africa: Council for Medical Schemes, available at http://www.medicalschemes.com.

Council for Medical Schemes (2006b). Annual Report 2006-07. Pretoria: Council for Medical Schemes, available at http://www.medicalschemes.com/

Koenker, R. and Hallock K. (2001). Quantile Regression: An Introduction. *Journal of Economic Perspectives*, 15:143–156.

Maurer, J. (2002). Modelling socioeconomic and health determinants of health-care use. *Health Econ*. 16: 976-979

McLeod H, Ramjee S. (2007). Medical Schemes. In: Harrison S, Bhana R, Ntuli A, editors. South African Health Review. http://www.hst.org.za/uploads/files/chap4_07.pdf

Republic of South Africa. Medical Schemes Act (Act 131of 1998). http://www.doh.gov.za/docs/legislation/acts/1998/act98-131.html

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics*,27:832-835.

Correspondence: m.willie@medicalschemes.com

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman & Hall.

Wilkinson, L. (1994). Less is More: Two- and Three-Dimensional Graphics for Data Display. *Behavior Research Methods, Instruments, & Computers*, 26:172-176.