Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 48-2011

Jing Dai, Walter Zucchini and Stefan Sperlich

Estimating and predicting the distribution of the number of visits to the medical doctor

This paper can be downloaded from
http://www.uni-marburg.de/fb02/makro/forschung/magkspapers/index_html%28magks%29

Coordination: Bernd Hayo • Philipps-University Marburg
Faculty of Business Administration and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Estimating and predicting the distribution of the number of visits to the medical doctor

Jing Dai, Universität Kassel,

Walter Zucchini, Universität Göttingen,

Stefan Sperlich Université de Genève,

November 21, 2011

Abstract

In many countries the demand for health care services is of increasing importance. Especially in the industrialized world with a changing demographic structure social insurances and politics face real challenges. Reliable predictors of those demand functions will therefore become invaluable tools. This article proposes a prediction method for the distribution of the number of visits to the medical doctor for a determined population, given a sample that is not necessarily taken from that population. It uses the estimated conditional sample distribution, and it can be applied for forecast scenarios. The methods are illustrated along data from Sidney. The introduced methodology can be applied as well to any other prediction problem of discrete distributions in real, future or any fictitious population. It is therefore also an excellent tool for future predictions, scenarios and policy evaluation.[1]

keywords: predicting health care demand, visits to the doctor, health economics, model selection
JEL classification: I12, C51, C53, H75.

[1]Corresponding author is Stefan Sperlich: Département des sciences économiques, Bd du Pont d'Arve 40, CH-1211 Genève 4, stefan.sperlich@unige.ch, tel: +41 22 3798223, fax: +41 22 3798229. This research was financially supported by the instituto de estudios fiscales, Madrid.

1 Introduction

A main challenge for health insurances is to analyze health demand[2]. One reason for this is the legitimate hope a better understanding will ease prediction, may it be for the future or new markets, that means in either case different populations. There, the frequency distribution of doctor consultations is a primary indicator of health care utilization in a population and, as such, is of obvious importance for health care budgeting. Therefore, patterns in the frequency of consultations to the doctor, especially the dependency of the utilization of health resources on demographic, socioeconomic and geographic factors has been extensively documented, and its proper modeling is of central interest for empirical research in health economics and applied econometrics, respectively (Cameron et al., 1988; Pohlmeier and Ulrich, 1995; Windmeijer and Santos Silva, 1997; Deb and Trivedi, 1997; Jochmann and León-González, 2004; Winkelmann, 2004; to mention only few). Typically, the literature bemoans a lack of data on certain key variables on the one hand and the 'shrinkage effect' effect of conditional expectations on the other hand. More recently, Berzel et al. (2006) offered a plausible description of the number of doctor visits by modeling its dependence on a very limited number of demographic factors. In fact, it turned out that the mean number of doctor visits can already be estimated quite well when applying appropriate statistical modeling on simple available demographic factors such as age, gender and location. These are excellent news as an important job for the future will be to tackle the increasing demand of health care services due to a drastically changing demography, especially in most parts of the industrialized world. In the Australian Capital Territory for example, where our data are taken from, the average age has increased from about 29 in 1990 to about 35 in 2010. The serious distributional change can be seen in Figure 1 from the Australian Bureau of Statistics:

The target is to predict the numbers of visits for a population having at hand only these simple demographic factors for the population of interest, but full information - i.e. also the numbers of visits - observed for a particular sample. There exist several well-studied methods for estimating missing values (see Dempster et al., 1977; Little and Rubin, 1987; Rubin, 1996; Schafer, 1997), some of which could be used in our context. Then, instead of estimating the distribution of interest directly one could consider the numbers of visits as

[2]See for example the contributions of to the special issue of *The Journal of Risk and Insurance 2010, Vol 77*

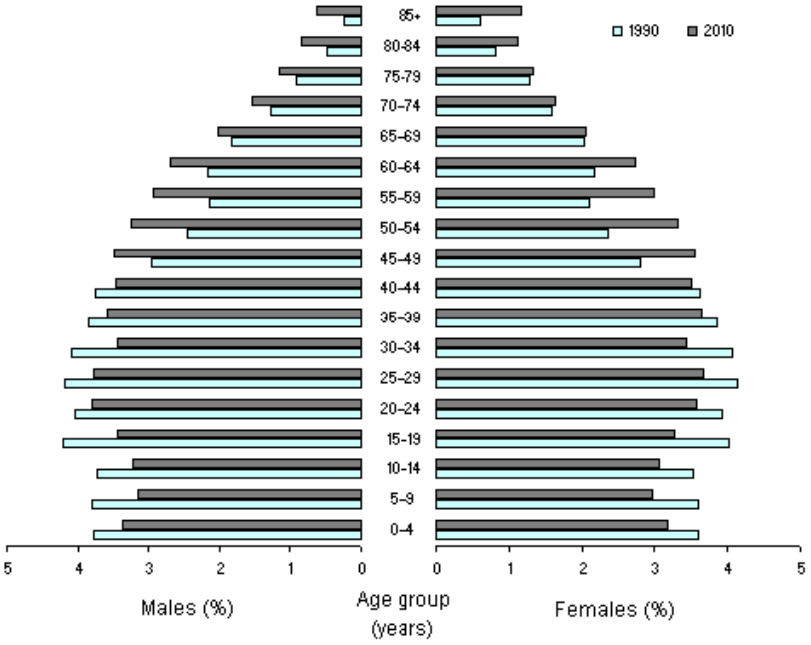

Figure 1: The population structure of Australia from 1990 to 2010 by age and sex. Source: The Australian Bureau of Statistics

missing values and complete the data by simulated (imputed) values. Most of these imputation processes are iterative. A key feature of such approaches is to regard missing data as random variables, and then to replace them with multiple draws from the assumed underlying distribution. Therefore, these methods are often known as 'multiple imputation'.

However, to our knowledge, no direct estimator of the population distribution of the doctor consultations has been reported yet. Note that the above-mentioned methods perform well for imputing missing values but have not been studied for imputing values for a whole population. Just thinking of the computational burden, if - as typically the case - the underlying sample is small but the population is large, these methods are not really attractive for our problem. Although the method we introduce here is straight forwardly applicable on many similar estimation or prediction problems, we concentrate on estimating the population distribution of doctor consultation frequencies, based on a moderate sample.

In a first step, Section 2, we search for a reasonable conditional distribution model based on only those covariates that are available in both the sample and the population of interest. As commonly acknowledged, the Poisson or negative binomial generalized linear model is the simplest way to model count data. The chosen link is typically the logarithm, i.e. the canonical link. According to the exploratory analysis, however, it may not be appropriate to use Poisson

or negative binomial generalized linear models for our data problem since the generalized linear model does not allow for the overdispersion parameter to depend on covariates. As a way of overcoming these limitations associated with Generalized Linear Models (GLM, see Nelder and Wedderburn, 1972) we tried also the generalized additive model for location, scale and shape (GAMLSS), introduced by Rigby and Stasinopoulos (2005). Nevertheless, all distribution models in question should be adapted to the sample, as the final objective is the optimal prediction or estimation of the unconditional distribution(s) of the population(s) of interest. In a second step, Section 3, one can now derive these distributions of interest as being a mixture of $N(=$*population size)* of the above calculated conditional distributions. All we need is a clear idea of the distribution of the covariates in the populations of interest and the assumption that the a priori fitted conditional models hold throughout. In Section 3 we present the numerical results. Section 4 concludes.

2 Modeling of the conditional distributions

The data set considered in this paper records the $23,607$ inhabitants of the Sydney suburb Ryde in 1994 and 1995. The available information comprises age, gender, and the number of doctor visits for both years. A more detailed description of these data can be found in Heller (1997). In the original data set there are 11 individuals (of the $23,618$) reporting more than 100 visits. As it turned out that this was due to an excessive misuse of the health insurance card by illegal immigrants for which is was impossible to obtain reliable corrections, we decided to truncate the data at a maximum of 100 visits. Note that then we have just 41 individuals in 1994, and 40 in 1995, with more than 52 visits, i.e. more than one each week. As no information is available that would allow for a sound detection of missmeasurement, we have not truncated these counts. The summary statistics for the remaining set of $N = 23,607$ inhabitants are given in Table 1.

There are mainly two prediction problems of interest. First, practitioners usually only have access to surveys, which for local areas can be of moderate size. From these they have to estimate the number of visits for a certain population, or to predict them for an artificial population to calculate scenarios. For example, in most industrialized countries a serious demographic change is expected in the next two decades which will effect the health systems and pension funds. In order to simulate these two situations we first draw a random

Table 1: Summary statistics, standard deviations in parentheses.

	population		sample	
	men	*women*	*men*	*women*
number of individuals	11302	12305	101	99
average age in 1994	36 (22)	39 (24)	37 (21)	40 (23)
average age in 1995	37 (22)	40 (24)	–	–
average number of visits in 1994	5.2 (6.3)	6.9 (7.2)	5.0 (5.3)	6.8 (6.1)
average number of visits in 1995	5.6 (6.5)	7.2 (7.2)	–	–

sample of only 200 observations from 1994 with the summary statistics given in Table 1. The extension to stratified sampling or other sampling schemes is obvious. The aim is to estimate the distribution of the number of visits to the medical doctor in 1994, and afterwards to predict it for 1995.

On the one hand it is well known that *gender* strongly interacts with age when looking at visits to the doctor; on the other hand, *age* is the only additional variable. Therefore, we first have to decide whether for a reasonable model fit the sample should be split by gender. In order to study this, we plot the number of doctor visits against age in Figure 2, separately for male and female. The solid and dotted lines are simple local linear regression estimates. They indicate a non-linear relationship between the mean of the number of doctor visits with age and gender. Furthermore, the differences between males and females seem to be quite complex and hard to capture in one common model.

Secondly, we analyze the variance-mean ratio to check for under or over dispersion. Figure 3 shows the variance-mean ratio by age and gender for the random sample of 200 inhabitants in 1994. The ratio is clearly greater than one for all levels of *age*, indicating inappropriateness of the Poisson model because of over dispersion. This exploratory analysis also reveals that *age* and *gender* have a strong influence on both the mean as well as the variance of visits; compare Figure 2 and 3.

Recall that the negative binomial regression model allows for overdispersion by introducing an unobserved heterogeneity term for each observation i. Observations are assumed to differ randomly in a manner that is not fully accounted for by the observed covariates. It assumes a negative binomial distribution for the response variable y in which its mean μ is modeled as a function of explanatory variables and a variance of the form $\mu + \mu^2\sigma$, where σ is an unknown overdispersion parameter which in turn shows no extra dependency on the covariate values. However, from Figure 3 we notice that the variance-mean ratio varies substantially over the covariate values. Consequently neither the

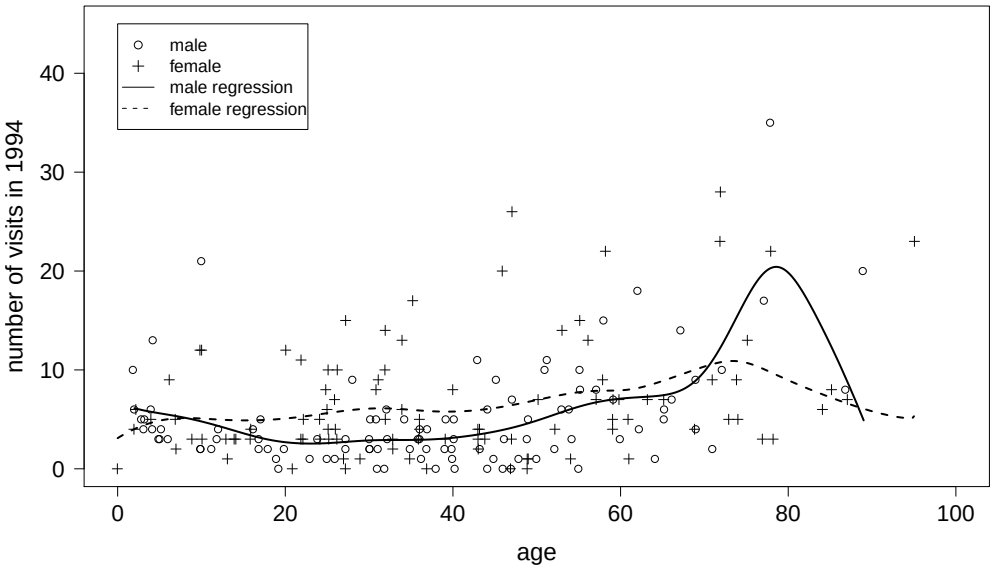

Figure 2: The number of visits to a GP (left, in 1994; right, in 1995) plotted against age for a simple random sample of 200 residents in Ryde. Local linear regression estimate with cross-validation bandwidth $\hat{h}_{CV} = 2.78$ (black line, male) and $\hat{h}_{CV} = 2.78$ (grey line, female)

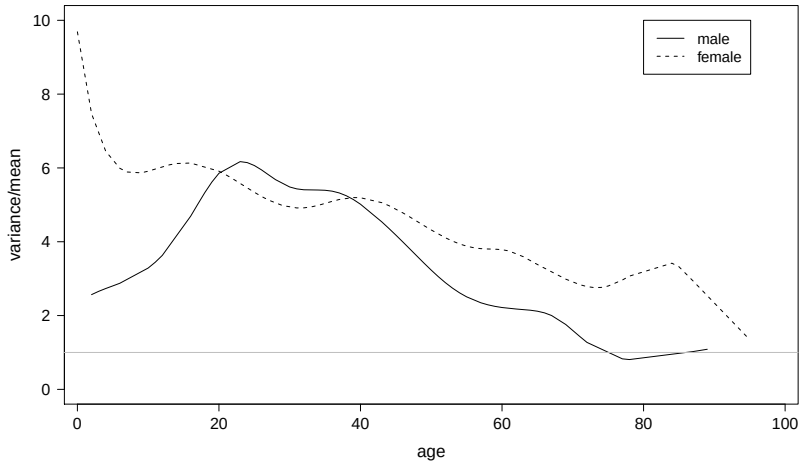

Figure 3: Variance by mean, separate for males and females, based on 200 random samples in 1994.

standard Poisson nor negative binomial generalized linear models seem to be appropriate in this case.

As indicated, we will need to fit appropriate models of conditional distributions to our data. Given our count data and the above findings we start with the negative binomial model (see for example, Cameron and Trivedi, 1998, Section

4.2.2), defined by

$$f(y|\mu,\sigma) = \begin{cases} \frac{\Gamma(y+1/\sigma)}{\Gamma(y+1)\Gamma(1/\sigma)} \frac{(\mu\sigma)^y}{(\mu\sigma+1)^{(y+(1/\sigma))}} & \text{if } x = 0, 1, \cdots \\ 0 & \text{otherwise} \end{cases}$$

with mean μ and variance $\mu + \mu^2\sigma$, see above. If the overdispersion is mainly due to zero inflations, an alternative extension of the simple Poisson is the zero inflated Poisson, i.e.

$$f(y|\mu,\alpha) = (1-\alpha)\cdot Po(y,0) + \alpha \cdot Po(y,\mu), \quad Po(y,\mu) = e^{-\mu}\mu^y/y!, \quad (2.1)$$

where again μ is modeled as a function of the covariates whereas α is an unknown scalar. An alternative to this extension of the Poisson we can also consider a zero inflated negative binomial having μ as a function of covariates and two unknown parameters σ and α. Different approaches to tackle the zero-inflation or other finite mixtures are proposed e.g. by Gurmu (1997), Deb and Trivedi (1997). See that issue also for further suggestions though in different contexts. As we mentioned before, for modeling linear functions, the linear models, lm(), and generalized lineal models, glm() of Hastie and Pregibon (1992) in the R language can be used. However we are restricted to model only the mean using lm() and glm().

In order to compare these three models we calculate the log-likelihood (llh), the deviance difference ΔD (relative to the simple Poisson) and the AIC of the fitted models as quality of fit statistics. The results are listed in Tables 2 and 3 respectively, separated by gender. Note first that the different criteria do not contradict each other. The zero-inflated Poisson model provides a slightly better fit than the Poisson model (not shown). However, the model which is superior (according to the AIC) is the negative binomial. The zero-inflated negative binomial shows no improvement compared to the negative binomial because the zero inflation is unnecessary after the inclusion of σ. Consequently, the observed deviance difference is zero relative to the negative binomial. The AIC even indicates that the improvement in fit is insufficient to justify the use of the more flexible but also more complex model. Recall that our main objective is not the optimal fitting but prediction, which is much more sensitive to overfitting due to complexity. Indeed, complexity is often one of the worst enemies of good prediction.

However, the generalized linear considered so far is restricted to allow only the location parameter to depend on covariates, and this only in a known para-

Table 2: Quality of fit statistics using GLM (for males)

Model	*Link*	*Terms*	*llh*	ΔD	*AIC*
zero-inflated Poisson	$log(\mu)$	$age + age^2$	-294	$-$	596
negative binomial	$log(\mu)$	$age + age^2$	-253	82	515
zero-inflated negative binomial	$log(\mu)$	$age + age^2$	-253	82	517

Table 3: Quality of fit statistics using GLM (for females)

Model	*Link*	*Terms*	*llh*	ΔD	*AIC*
zero-inflated Poisson	$log(\mu)$	$age + age^2$	-360	$-$	728
negative binomial	$log(\mu)$	$age + age^2$	-286	148	579
zero-inflated negative binomial	$log(\mu)$	$age + age^2$	-286	148	581

metric way. Rigby and Stasinopoulos (1996, 2005) developed a general class of univariate regression models, called the Generalized Additive Model for Location, Scale and Shape (GAMLSS) with two important extensions. First, they allow all distribution parameters to depend on a predetermined set of covariates. Second, the modeling of these parameter functions may include random effects or even be nonparametric, but being always of an additive structure. The model assumes independent observations of the response variable given the parameters, the covariates and the values of the random effects. It provides a very general distribution family for univariate continuous or discrete response variables. In our case, under the negative binomial distributional assumption, both the mean and the dispersion parameter can be modeled as a function of *age*. To summarize, we consider the negative binominal density $f(y|\mu, \sigma)$ and will estimate

$$\log(\mu) = g_1(age), \qquad \log(\sigma) = g_2(age), \tag{2.2}$$

where we first will set g_1, g_2 to be parametric quadratic function, and afterwards nonparametric cubic splines (cs). For the latter we have plotted the functions g_1, g_2 in Figure 4.

For comparing these two GAMLSS models, we use the well known fitted global deviances $GD = -2l(\hat{\theta}) = -2\sum_{i=1}^{n} l(\hat{\theta}^*)$, the Akaike information criterion AIC of Akaike (1974) and the Schwarz Bayesian criterion SBC of Schwarz (1978). AIC and SBC are asymptotically justified as predicting the degree of fit in a new data set, i.e. approximations to the average predictive error. The global deviance, SBC and AIC are summarized as statistics relating to the fit of the parametric and nonparametric GAMLSS models in Table 4 and 5, again separately for males and females. Fortunately, the different criteria do

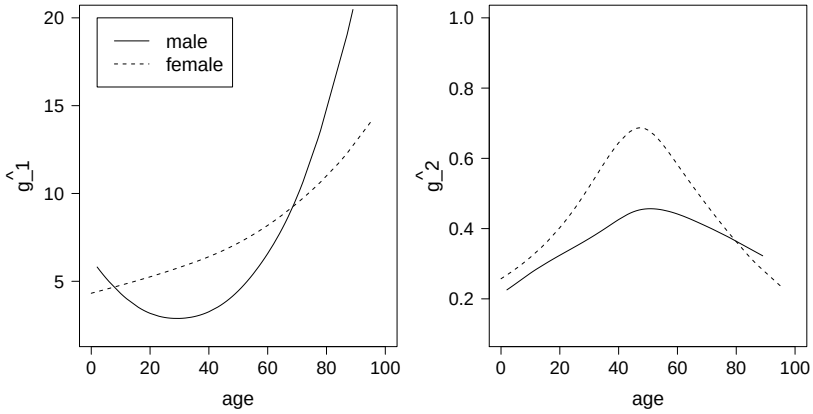

Figure 4: Impact of age and gender on the GAMLSS nonparametric regression estimates for mean g_1 (left) and dispersion g_2 (right), based on a random sample of 200 residents in Ryde in 1994.

Table 4: Quality of fit statistics using GAMLSS (for males)

Model	*Link*	*terms*	*GD*	*AIC*	*SBC*
negative binomial	$log(\mu)$	$age + age^2$	506	518	533
(parametric model)	$log(\sigma)$	$age + age^2$			
negative binomial	$log(\mu)$	$cs(age)$	502	515	534
(nonparametric model)	$log(\sigma)$	$cs(age)$			

give the same selections so that it is enough to look at the AIC here.

A further possibility to model dispersion in parametric or nonparametric negative binomial regression is the Vector Generalized Additive Model introduced by Yee and Wild (1996). One can also find some discussions about applying the provided R-package VGAM for count data in Berzel et al.(2006). However, already now we can see, compare Tables 2 to 5 that the AIC always selects the negative binomial generalized linear model throughout. This confirms our statement that, depending on the amount of information (data and signal-noise ratio), complexity is one of the worst enemies of prediction. Consequently, it is questionable to what extend other flexible, semi- or non-parametric model approaches can improve in our prediction problem. Nevertheless, in the final step we will also consider the GAMLSS results for the following reason. Our

Table 5: Quality of fit statistics using GAMLSS (for females)

Model	*Link*	*terms*	*GD*	*AIC*	*SBC*
negative binomial	$log(\mu)$	$age + age^2$	568	580	596
(parametric model)	$log(\sigma)$	$age + age^2$			
negative binomial	$log(\mu)$	$cs(age)$	568	580	595
(nonparametric model)	$log(\sigma)$	$cs(age)$			

objective is not the conditional but the unconditional density of visits, and we do not know which model yields the best results there. Figure 4 shows that the data fit indicates a nonlinear, nonconstant dispersion parameter. While limiting to a quadratic modeling seems adequate, ignoring this finding might cause prediction loss in the final step.

3 Predicting the population distribution

In this section, we come to this final step when applying our new method to the two described problems. Both are for prediction: ones is using a random sample of 101 males, and 99 females respectively, in 1994 to estimate the distribution of number of visits for the male (and female) population in Ryde in the same year (case study 1); another one is to predict the number of visits prediction for male (or/and female) population in Ryde in 1995 using the same random sample of males (or/and females) in 1994 (case study 2).

3.1 Case study 1

We start with estimating the distribution of number of doctor visits for the population, given a sample in the same year. The method we suggest is not simulation based; it provides transparent and reproducible estimators. We have a sample $\{(x_i^s, y_i^s)\}_{i=1}^n$, and the covariates $\{x_j\}_{j=1}^N$ of the population of interest. Recall that the required unconditional distribution of the population of interest, say $f_N(y)$, is simply the marginal distribution of the joint density $f_N(y, x)$ such that

$$f_N(y) = \int f_N(y, x)dx = \int f_N(y|x)f_N(x)dx \ . \tag{3.1}$$

For finite populations we can simplify to

$$f_N(x) = \begin{cases} \frac{1}{N} & \text{if } x = x_j \\ 0 & \text{if } x \neq x_j \end{cases}$$

and then obtain

$$f_N(y) = \frac{1}{N}\sum_{j=1}^{N} f_N(y|x_j). \tag{3.2}$$

Thus, what we need is a reasonable substitute in equation (3.2) for the conditional densities $f_N(y|x)$. An obvious choice here is one of the conditional densities fitted to the sample data, say $f_n(y|x)$. If $f_n(y|x)$ is a consistent estimate of $f_N(y|x)$, the consistency for $f_N(y)$ follows immediately. Also the asymptotic properties can be derived directly for most cases via Taylor expansion. In the nonparametric case this can be quite tedious, compare e.g. Van Keilegom and Veraverbeke (2002) or Sperlich (2009). In both the nonparametric and the parametric world, the estimator of the unconditional density will inherit consistency and convergence rate from the conditional density estimate.

What happens if $f_n(y|x)$ is not a consistent estimate of $f_N(y|x)$? In that case our procedure will still give a good approximate for $f_N(y)$ as long as the relation between y and covariates x specified and estimated from the sample can be carried over to the population reasonably well. In that case one could think of

$$\hat{f}_N = \frac{1}{N}\sum_{j=1}^{N} f_n(y|x_j)$$

as an N-fold mixture of pre-determined densities relating y to some covariates x.

In our case we considered the negative binomial (NB) as a reasonable description of the relation between y and x. With the estimates $\hat{g}_1, \hat{g}_2$ obtained from our sample $\{(x_i^s, y_i^s)\}_{i=1}^n$ in Section 2 we estimate then the unconditional probability function of y by

$$\hat{f}_N(y) = \frac{1}{N}\sum_{j=1}^{N} NB[y|\hat{g}_1(x_j), \hat{g}_2(x_j)]. \tag{3.3}$$

For the different specifications, NB with quadratic g_1 and constant g_2 (the GLM), g_1 and g_2 quadratic functions of age, and finally cubic splines for g_1 and g_2 (GAMLSS), the results are given in Figure 5. As in our example we have records of the real number of visits, we can evaluate our predictions exactly. It can be seen that the distribution of the estimated conditional mean (circles) is much too narrow to be of use when we are interested in the distribution of real visits (bars). In contrast, the predicted unconditional distribution (solid circles) fits very well. While for men, it seems that it would be worth to pay more attention on a possible zero inflation, the problem is less emphasized for females.

For comparing the different GLM and GAMLSS specifications we need a more

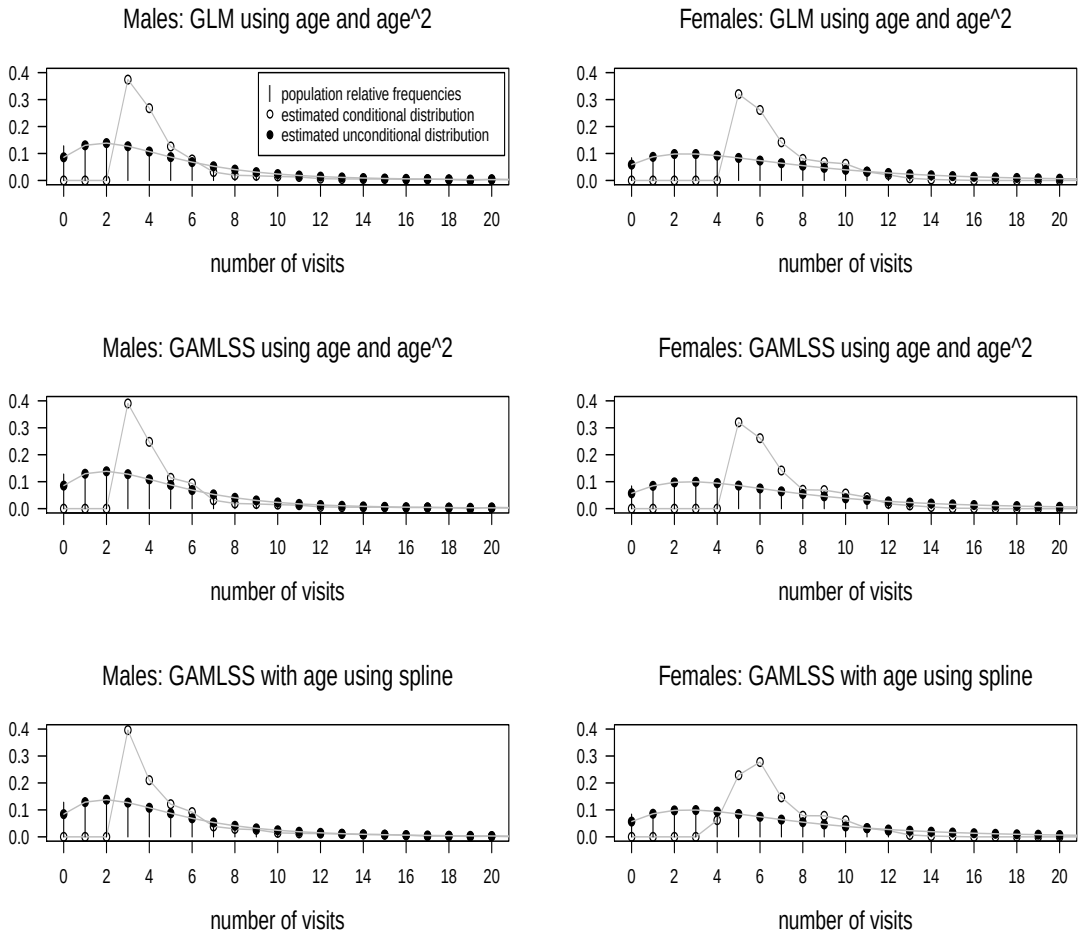

Figure 5: Predicted population distribution based on negative binomial GLM estimates (upper), negative binomial GAMLSS parametric (middle) and cubic spline (lower) specification for 1994.

careful analysis. In order to do so, we calculated the prediction error

$$\frac{1}{M}\sum_{m=1}^{M} LOSS[\hat{f}_N(y_m) - f_N(y_m)], \tag{3.4}$$

where M is the number of values y does take, i.e. $1, 2, \cdots, 42$ for males and $1, 2, \cdots, 34$ for females. $LOSS[\cdot]$ stands simply for $abs[\cdot]$ (L1-norm) and $[\cdot]^2$ (L2-norm) respectively. The outcome is listed in Table 6. According to this, the negative binomial GAMLSS using spline performs best for both males and females. It might however be surprising that for males it does much better than GLM although the AIC was the same (515 for both CS-GAMLSS and GLM). The problem with the AIC is that one needs to calculate the degrees of freedom which can be quite problematic in nonparametric statistics, see e.g. Sperlich et al. (1999) or Müller (2001). Note finally that based on our observation in Figure 5 we also calculated the prediction errors for the zero-inflated NB GLM; surprisingly, it never outperforms the NB GAMLSS using

Table 6: L1 and L2-Norm prediction errors of case 1

	L1-Norm		L2-Norm	
Model	*males*	*females*	*males*	*females*
negative binomial GLM	.02480	.03558	.00385	.00451
zero-inflated negative binomial	.02481	.03558	.00390	.00451
negative binomial GAMLSS	.02483	.03531	.00389	.00448
NB GAMLSS using spline	.02334	.03193	.00371	.00346

spline.

3.2 Case study 2

Even more challenging - and also more interesting for health economics and political decision making - is the prediction of visits to the medical doctor for the future.

Clearly, the theoretical findings from equations (3.1) to (3.3) stay all the same. The only difference is that, at least for far horizons, it is to be expected that the relation between y and the used covariates will change. The prediction performance of our method to the future depends on the persistency of the relation we estimate from the sample. To keep the problem simple we will use the same sample, i.e. the results obtained in Section 2 to now predict the distribution of visits to the doctor for 1995. Applying the same procedure as we used in the first case study, we get the predictions illustrated in Figure 6. At first glance the prediction performance looks even better than the estimation performance in case 1. This is due to the lack of zero inflations in the recorded real visits. This now also explains why the different criteria in Section 2 opted for models without zero-inflation. All these criteria are constructed thinking of an infinite hyperpopulation, i.e. of a distribution from which the populations in 1994 and 1995 are just random samples. Then, a zero-inflation would fit better the 1994 data but constitutes an overfit for the hyperpopulation.

As for the first case study we again analyzed the prediction errors, see equation (3.4), of our different specifications, summarized in Table (7). We get a similar ranking of the specifications as in case study 1 but, as already noted from Figure 6, with better total performance. Again, the nonparametric NB GAMLSS clearly gives the best predictions for both the male and female populations.

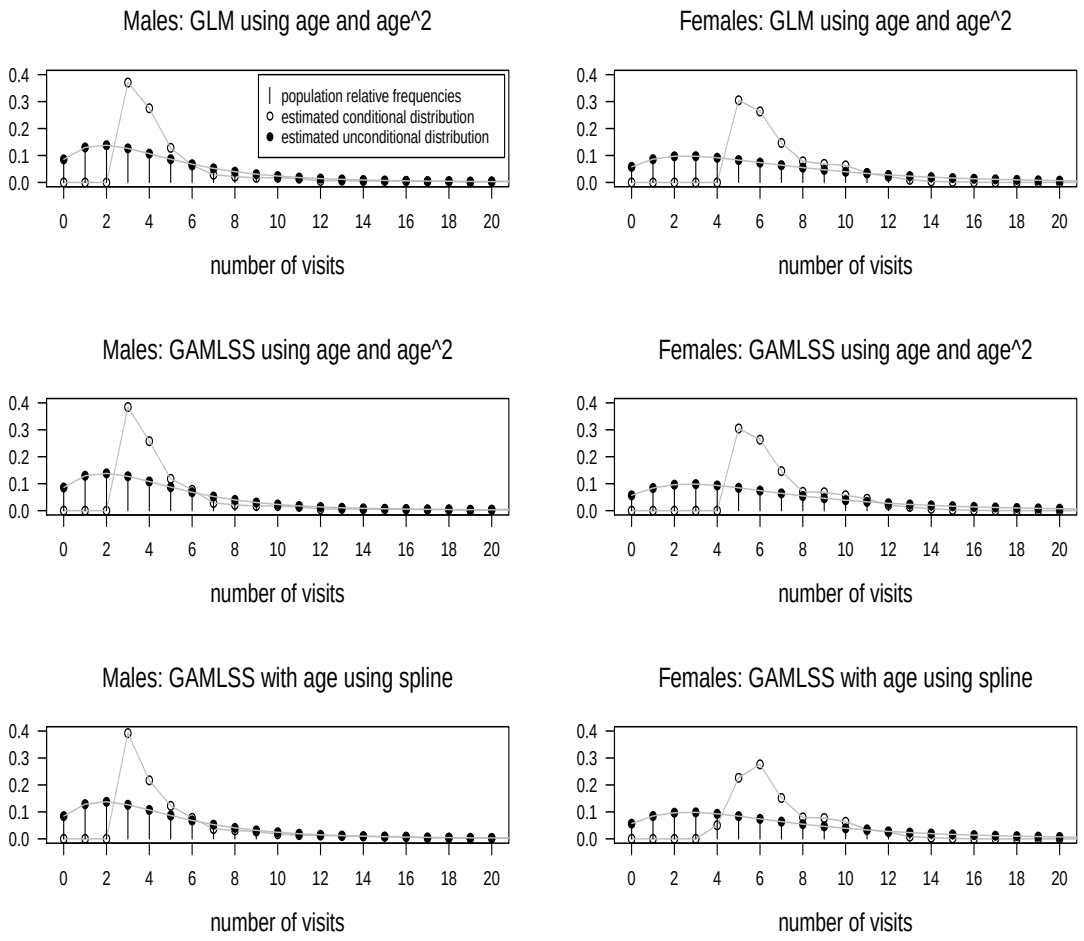

Figure 6: Predicted population distribution based on negative binomial GLM estimates (upper), negative binomial GAMLSS parametric (middle) and cubic spline (lower) specification for 1995.

Table 7: L1 and L2-Norm prediction errors of case 2

	L1-Norm		L2-Norm	
Model	*males*	*females*	*males*	*females*
negative binomial GLM	.02402	.03407	.00367	.00411
zero-inflated negative binomial	.02403	.03407	.00371	.00411
negative binomial GAMLSS	.02405	.03369	.00368	.00409
NB GAMLSS using spline	.02238	.03110	.00348	.00325

4 Conclusions

The number of visits to the medical doctor is of essential interest in health economics, be it for political decision making or the planning of health care institutions and insurance. While there exist many econometric model approaches to study the demand for health care, it is hard to find a simple but effective method to estimate and predict the unconditional distribution of the

number of visits. Often demographic information which is easily available even on the census level turns out to be more helpful for estimating – not to mention for prediction and scenario simulations – the distribution of the number of visits than complex econometric models. Doubtless the latter have their particular justification in more sophisticated analysis.

A careful model specification and selection is the necessary prior step to obtain the conditional relationship between the number of visits and the demographic factors. Here, we fitted different conditional densities to the sample data. Then a well known integration principle yields a predictor for the required unconditional distribution of visits. In case the conditional sample distribution is a consistent estimator for the population analogue, this predictor inherits its asymptotic properties, in particular consistency and convergence rate. In case the population of interest follows a different distribution than the sample at hand, we still have the interpretation of our predictor as an intuitively appropriate N-fold mixture distribution. This may explain the excellent performance of our method despite its simplicity, in both case studies: for the estimation of the unconditional population distribution, and for the future prediction.

The model selection might be done via standard criteria as we used. However, one should not forget that these try to select the best estimator for the conditional distribution whereas the final objective is the unconditional one. This or the problematic calculation of the degrees of freedom for nonparametric estimators would explain that, for example, the AIC did not select the optimal model for the prediction of female visits to the medical doctor. In our case studies we were in the fortunate situation of knowing the outcome and could therefore compare the prediction with the real number of visits. In practice we recommend to evaluate the final predictor of the unconditional distribution on the observed sample. Note that due to the explicit analytic form of our estimator / predictor our results are transparent and reproducible. We do not use any random-, simulation- or resampling methods. As a flexible method it can be incorporated in any mixed effects or more sophisticated econometric models. Furthermore, it can be easily extended to any other context and allows for further inference.

Moreover, the here originally introduced method can be applied equally well to any other prediction problem of discrete distributions. It does not matter whether these are predictions (or estimations) for presently real, possible future or any fictitious population. It is therefore also an excellent tool for future predictions, scenarios and policy evaluation.

References

Akaike, H. 1974. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**: 117-128.

Berzel, A., Heller, G.Z. and Zucchini, W. 2006. Estimating the number of visits to the doctor. *Australian & New Zealand Journal of Statistics* **48**: 213-224.

Cameron, A.C. and Trivedi, P.K. 1998. *Regression Analysis of Count Data.* Cambridge: Cambridge University Press.

Cameron, A.C., Trivedi, P.K., Milne, F. and Piggott, J. 1988. A microeconometric model of the demand for health care and health insurance in Australia. *Review of Economic Studies* **55**: 85-106.

Deb, P. and Trivedi P.K. 1997. Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* **12**: 313-336.

Gutmu, S. 1997. Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. *Journal of Applied Econometrics* **12**: 225-242.

Hastie, T. J. and Pregibon, D. 1992. *Genaralized Linear Model.* Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Heller, G.Z. 1997. Who visits the GP? Demographic patterns in a Sydney suburb. *Technical report, Department of Statistics, Macquarie University.*

Dempster, A.P., Laird, N.M. and Rubin, D.B. 1977. Maximum-Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**: 1-38.

Jochmann, M. and León-González, R. 2004. Estimating the demand for health care with panel data: a semiparametric Bayesian approach. *Health Economics* **13**: 1003-1014.

Little, R.J.A. and Rubin, D.B. 1987. *Statistical Analysis with Missing Data.* New York, John Wiley.

Müller, M. 2001. Estimation and testing in generalized partial linear models – A comparative study. *Statistics and Computing* **11**: 299-309.

Nelder, J.A. and Wedderburn, R.W.M. 1972. Generalized linear models. *Journal of the Royal Statistical Society A* **135**: 370-384.

Pohlmeier, W. and Ulrich, V. 1995. An econometric model of the two-part decision making process in the demand for health care. *Journal of Human Resources* **30**: 339-361.

Rigby, R.A. and Stasinopoulos, D.M. 1996. A Semi-parametric Additive Model for Variance Heterogeneity. *Statistical Computing* **6**: 57-65.

Rigby, R.A. and Stasinopoulos, D.M. 2005. Generalized additive models for location, scale and shape. *Applied Statistics* **54**: 507-554.

Rubin, D.B. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**: 473-489.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* **6**: 461-464.

Sperlich, S. 2009. A note on non-parametric estimation with predicted variables. *The Econometrics Journal* **12**: 382-395.

Sperlich, S., Linton, O., and Härdle, W. 1999. Integration and backfitting methods in additive models. – Finite sample properties and comparison. *Test* **8**: 419-458.

Van Keilegom, I. and Veraverbeke, N. 2002. Density and hazard estimation in censored regression models. *Bernoulli* **8**: 607-625.

Windmeijer, F.A.G. and Santos Silva J.M.C. 1997. Endogeneity in count data models: a application to demand for health care. *Journal of Applied Econometrics* **12**: 281-294.

Winkelmann, R. 2004. Health care reform and the number of doctor visits – an econometric analysis. *Journal of Applied Econometrics* **19**: 455-472.

Yee, T. W. and Wild, C. J. 1996. Vector generalized additive models. *Journal of Royal Statistical Society, Series B, Methodological* **58**: 481-493.