HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 16/20

Semiparametric Count Data Modeling with an Application to Health Service Demand

Philipp Bach, Helmut Farbmacher & Martin Spindler

August 2016

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

Semiparametric Count Data Modeling with an Application to Health Service Demand

Philipp Bach*, Helmut Farbmacher[b], Martin Spindler[a,b]

[a]Hamburg Business School
University of Hamburg

[b]Center for the Economics of Aging
Max Planck Society, Munich

*bach.philipp@outlook.com

July 18, 2016

SUMMARY

Heterogeneous effects are prevalent in many economic settings. As the functional form between outcomes and regressors is often unknown a-priori, we propose a semiparametric negative binomial count data model based on the local likelihood approach and generalized product kernels, and apply the estimator to model demand for health care. The local likelihood framework allows us to leave the functional form of the conditional mean unspecified while still exploiting basic assumptions in the count data literature (e.g., non-negativity). The generalized product kernels allow us to simultaneously model discrete and continuous regressors, which reduces the curse of dimensionality and increases its applicability as many regressors in the demand model for health care are discrete.

JEL codes: I10, C14, C25
Keywords: semiparametric; nonparametric; count data; health care demand

1 INTRODUCTION

Estimating the demand for health services is a major field of application of count data regression since frequently observed outcome variables of interest only assume non-negative integer values, for instance, the number of doctor visits or hospitals stays. Studies in this discipline of health economics aim at assessing the impact of health-related, socio-economic or insurance-related characteristics on individuals' health care demand. The predominant regression techniques for modeling the health service demand are entirely parametric, for example the Poisson, negative binomial regression models, zero-inflated models and hurdle models. Being typically estimated by the method of maximum likelihood, these models basically incorporate two types of assumptions: First, assumptions on the distribution of the outcome variable and, second, the parametrization of the conditional mean. The latter specifies the relationship of the outcome variable and the regressors which is assumed to be log-linear, in order to account for the non-negativity of the response, i.e. $\mathrm{E}[y|x] = \exp(x'\beta)$.[1] While a considerable number of count data models allow to abstract from distributional assumptions on y, consistency of almost all parametric and semiparametric estimators hinges on the correct specification of the conditional mean $\mathrm{E}[y|x]$.

Winkelmann (2008) points out, that the log-linearity assumption on the conditional mean might be violated frequently. In the context of health service demand, nonlinearities might arise due to heterogeneous effects of certain characteristics on therapeutic treatment intensity. One can think of a differential effect of the number and type of chronic conditions. For example, a change from zero to one chronic condition might have a different effect on the number of doctor visits than a change from three to four chronic diseases or, alternatively, diabetes might have a different effect on the number of hospital stays than cardiovascular conditions. It is also possible to think of a nonlinear effect of age referring to age-related diseases and interaction effects, for instance with gender, occupation or income. Moreover, it can be thought of a heterogeneity of the effect of insurance status that might be different for low-, median- and high-income individuals. Given this argumentation, defenders of parametric count data regression might counter that the log-linearity assumption serves as an approximation and that nonlinearities might enter the models by inclusion of interaction or higher polynomial terms. However, in many settings it is difficult to specify the relationship of the outcome variables and the regressors *ex ante* due to highly complex economic, psychological or biological processes.

[1] Count data models with a more complex structure, e.g. finite mixture models, hurdle models and zero-inflated models, embody slight variations of the conditional mean assumption due to an incorporated weighting associated with multiple classes (Cameron and Trivedi, 2013).

And even if a log-linear conditional mean was considered an approximation to the true relationship of the regressors and the outcome variable, studies that rely on parametric regression techniques might miss to reveal heterogeneous patterns in the data (Winkelmann, 2008; McLeod, 2011). Indeed, the count data models incorporating the assumption $\mathrm{E}[y|x] = \exp(x'\beta)$ do neither allow the regression coefficients to change signs nor to vary at all across individuals (Frölich, 2006).

In this paper we propose a semiparametric negative binomial type 2 estimator that is based on the local likelihood approach. It allows us to to abstract from the log-linearity assumption on the conditional mean and to account for overdispersion at the same time. The local likelihood approach as initially developed by Tibshirani and Hastie (1987) is introduced to the context of modeling health service demand. Local likelihood estimation is a well-studied method in the statistics literature (Tibshirani and Hastie, 1987; Fan *et al.*, 1995; Fan and Gijbels, 1996; Fan *et al.*, 1998) and has only recently been introduced to the context of count data regression by Santos and Neves (2008). Basically, the local likelihood approach is appealing for two reasons: First, it is sufficiently flexible in order to leave the relationship of the covariates and the conditional mean of the independent variable unspecified, and, thus, allows for potential nonlinearities. Second, it maintains a likelihood structure and, hence, allows to develop specific estimators for count data regression and to achieve efficiency gains compared to fully nonparametric estimators (Frölich, 2006).

This paper contributes to the literature on count data regression in various aspects. It extends the first study on local likelihood estimation in the context of count data regression by Santos and Neves (2008) to settings with *mixed data*, i.e. the set of regressors includes categorical and continuous variables - a situation frequently encountered in estimating the demand for health services (Jones *et al.*, 2013). For instance, dummies for gender or categorical variables for health status are regularly included in the corresponding regression models. Moreover, in contrast to the semiparametric Poisson estimator by Santos and Neves (2008), the local likelihood negative binomial type 2 estimator derived in this paper is compatible with overdispersed data. Furthermore, the paper offers the first goodness-of-fit comparison of a local likelihood estimator for count data regression to commonly implemented fully parametric and nonparametric estimators in a simulation study and an empirical application.

So far, only few methods exist that allow to abstract from the log-linearity assumption on the conditional mean whereas many of them focus on the choice of the exponential function as a response function, for instance Weisberg and Welsh (1994). So-called generalized partially linear models (Robinson, 1988) address the violation of a log-linear conditional mean function. In these models,

it is assumed that the log-linearity assumption holds for a part of the regressors, while it is known to be violated and hence left unspecified for the remaining fraction of explanatory variables. For instance, Severini and Staniswalis (1994) propose to estimate the unknown relationship by kernel weighted log-likelihood (Staniswalis, 1989). However, this approach is limited by the necessity of separating the covariates for which a log-linear relationship is known from those with an unspecified relation. Furthermore, the nonparametric part of the model only allows for continuous regressors (Cameron and Trivedi, 2013).

Due to the encountered limitations, the existing semiparametric methods may be of limited use in health economic settings. Alternatively, researchers might employ fully nonparametric methods that do not impose any assumptions on the relationship of the dependent variable and the regressors. In a recent study, McLeod (2011) applies a nonparametric kernel density estimator in order to model health service demand and finds a superior model fit compared to a finite mixture negative binomial type 2 model. Overall, fully nonparametric methods can be judged as non-specific in that they are generally applicable to any context and not explicitly developed for count data regression. Accordingly, they do not take the structure of the count variable as a non-negative integer into account. By incorporating a reasonable assumption on the error distribution in the count data model (i.e. non-negativity), the local likelihood approach allows to achieve efficiency gains as compared to fully nonparametric methods (Frölich, 2006).

The remainder of this paper is organized as follows: In the next section we derive our semiparametric local likelihood negative binomial type 2 estimator and extend the local likelihood framework to discrete regressors. In Section 3 we compare our model to fully parametric and nonparametric estimators in a simulation study. In Section 4 we illustrate the relevance of our estimator using a real-data empirical example. Section 5 concludes.

2 MODEL AND ESTIMATION

2.1 A local likelihood estimator for count data

Extending the work by Santos and Neves (2008), a local likelihood negative binomial type 2 (NB2) estimator is derived as a semiparametric estimator for count data regression which is compatible with (i) overdispersed and (ii) mixed data. A sample of n i.i.d. observations with outcome variable y_i and covariates x_i is considered. In line with the previous section, y_i is a count variable, i.e. it only assumes non-negative integer values, $y_i = 0, 1, 2, ...$ The data are *mixed*, i.e. the k independent variables are either *continuous* or *discrete* in

nature $x_i = (x_i^c, x_i^d)$. There are k_d discrete or, alternatively, *categorical*, and k_c continuous regressors, such that $k_d + k_c = k$. A categorical variable x_{is}^d, i.e. s-th component of the discrete regressors vector x_{is}^d, takes c_s different values with $c_s \geq 2$, i.e. $x_{is}^d \in \{0, 1, 2, ..., c_s - 1\}$, $s = 1, ..., k_d$.

The following presentation of the local likelihood NB2 estimator parallels that of the parametric benchmark model as depicted in Winkelmann (2008). In contrast to the parametric NB2 framework, the assumption of a log-linear conditional mean, $\mathrm{E}[y|x] = \mu = \exp(x'\beta)$ is dropped. The conditional probability function of y, $f(y|\mu, \sigma^2)$, is the negative binomial probability function

$$f(y|\mu, \sigma^2) = \frac{\Gamma(\sigma^{-2} + y)}{\Gamma(\sigma^{-2})\Gamma(y+1)} \left(\frac{\sigma^{-2}}{\mu + \sigma^{-2}}\right)^{\sigma^{-2}} \left(\frac{\mu}{\mu + \sigma^{-2}}\right)^{y} \tag{1}$$

where $E[y|x] = \mu$ and $\mathrm{Var}(y|x) = \mu + \sigma^2\mu^2$ denote the conditional mean and variance of the outcome variable of y with precision parameter σ^2.

The intuition of the local likelihood approach can be illustrated by a comparison of the semiparametric model setup to the parametric framework. In a parametric NB2 model, one would now specify the conditional mean $\mu = \exp(x'\beta)$ and maximize w.r.t. β accordingly. However, in the local likelihood framework we do not assume that the relationship $\mathrm{E}[y|x] = \exp(x'\beta)$ is known (to be log-linear in x). Instead, the relationship m in $\mu = \exp(m)$ is left unspecified and m is fitted locally by using a Taylor series approximation of degree p, m_p. In order to weight observations that are close to a certain point (y_i, x_i) more heavily, a kernel weighting K_γ is introduced to the log-likelihood function. The conditional locally weighted log-likelihood function is set up as:

$$\begin{aligned} \mathcal{L}_0(\mu_0, \sigma_0^2) \;=\; & \sum_{i=1}^{n} \Bigg[\Bigg\{ \left(\sum_{j=1}^{y_i} \log(\sigma_0^{-2} + j - 1) \right) - \log y_i! \\ & -(y_i + \sigma_0^{-2})\log(1 + \sigma_0^2\mu_0) + y_i \log\sigma_0^2 + y_i \log\mu_0 \Bigg\} K_{\gamma,i} \Bigg] \end{aligned} \tag{2}$$

Corresponding to the notation in Santos and Neves (2008), the subscript in $\mathcal{L}_0$ indicates that we use a local constant approximation for the unknown parameters, i.e. $\mu(x_0) \approx \mu_0$ and $\sigma^2(x_0) \approx \sigma_0^2$.[2] $\gamma = (h, \lambda)$ in (2) denotes

[2] Instead of imposing $\mu = exp(x'\beta)$, the $x'\beta$ is approximated locally by a Taylor series expansion. While in general, this approximation incorporates a polynomial of degree p, here only local constant approximations ($p = 0$) are treated for sake of notational simplicity. A more general presentation with p-th order polynomials can be found in Fan *et al.* (1995) or Fan *et al.* (1998).

the vector of smoothing parameters for the continuous (h) and discrete (λ) regressors.
The first-order conditions w.r.t. μ_0 and σ_0^2 then define the local constant estimators on (μ, σ^2):

$$\begin{aligned}
\frac{\partial \mathcal{L}_0}{\partial \mu_0} &= \sum_{i=1}^{n} \left\{ \frac{y_i}{\mu_0} - \frac{(y_i + \sigma_0^{-2})\sigma_0^2}{1 + \sigma_0^2 \mu_0} \right\} K_{\gamma,i} = 0 \\
\frac{\partial \mathcal{L}_0}{\partial \sigma_0^2} &= \sum_{i=1}^{n} \left\{ \frac{1}{\sigma_0^4} \left(\log(1 + \sigma_0^2 \mu_0) - \sum_{j=1}^{y_i} \frac{1}{\sigma_0^{-2} + j - 1} \right) \right. \\
&\quad \left. - \frac{(y + \sigma_0^{-2})\mu_0}{1 + \sigma_0^2 \mu_0} - \frac{y_i}{\sigma_0^2} \right\} K_{\gamma,i} = 0 \qquad (3)
\end{aligned}$$

From the FOC w.r.t. μ, one can derive the expression for the local constant estimator $\hat{\mu}_0$:

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} y_i K_{\gamma,i}}{\sum_{i=1}^{n} K_{\gamma,i}} \qquad (4)$$

Here, the result of Fan *et al.* (1995) on local likelihood estimation of generalized linear models (GLMs) can be verified, i.e. the expression for the local likelihood NB2 estimator coincides with that of the Nadaraya-Watson estimator. Accordingly, the local constant likelihood NB2 estimator is consistent under the minimum assumptions that are sufficient for consistency of the Nadaraya-Watson estimator (Li and Racine, 2007). It can be shown that the negative binomial 2 distribution with known ancillary parameter σ^{-2} belongs to the linear exponential family and hence the NB2 model can be classified as a GLM (Hilbe, 2011). The estimator $\hat{\sigma}_0^2$ can be obtained by using appropriate numerical methods. The asymptotic theory for local likelihood estimators in the context of GLMs can be found in Fan *et al.* (1995).

2.2 Kernels for continuous and discrete regressors

In order to develop a local likelihood estimator suitable for mixed data (i.e., discrete and continuous regressors), it is necessary to use kernel functions that account for the discrete nature of the regressors. We extend the local likelihood estimation to smooth over discrete variables which greatly extends the scope of applicability as in many economic applications the variable of interest is discrete (for instance, insurance status or treatment evaluations in general).

Building upon the work by Li and Racine (2007), so-called "*generalized product kernels*" are discussed in the following. The main advantage of using these kernel functions is that we can use all observations in the semi- and

nonparametric estimation instead of fitting the data separately for all possible combinations of the discrete regressors. Therefore the curse of dimensionality only comprises the continuous variables and, thus, is substantially less severe than in early versions of kernel regression when the so-called "frequency approach" was used.[3]

Paralleling Li and Racine (2007, 136), we define the kernel estimators for the continuous and discrete regressors separately. Note that in this section, a potential natural ordering of the independent variables is ignored. An extension to ordered regressors is straightforward by inserting an appropriate kernel function (Li and Racine, 2007).
For the continuous regressors, the product kernel $C_h(x^c, x_i^c)$ at a point $x = (x^c, x^d)$ with continuous part $x^c \equiv (x_1^c, ..., x_{k_c}^c)'$ is defined as

$$C_h(x^c, x_i^c) = \prod_{q=1}^{k_c} h_q^{-1} w_c \left(\frac{x_q^c - x_{iq}^c}{h_q} \right) , \tag{5}$$

where $h_q \in (0, \infty)$ is the bandwidth or smoothing parameter for regressor x_q^c, $q = 1, ..., k_c$, and w_c is a kernel function for the continuous regressors that is symmetric, nonnegative, univariate and satisfies standard assumptions listed in Li and Racine (2007, 9). In the Monte Carlo simulation and the application we use a Gaussian kernel.
For the discrete regressors x_s^d, with $s = 1, ..., k_d$, we define a product kernel function that incorporates a variation of the kernel function of Aitchison and Aitken (1976) such that[4]

$$w_{d,s}(x_s^d, x_{is}^d, \lambda_s) = \begin{cases} 1, & \text{if } x_{is}^d = x_s^d \\ \lambda_s, & \text{otherwise} \end{cases} \tag{6}$$

with smoothing parameter $\lambda_s \in [0, 1]$. Accordingly, the product kernel for the discrete regressors becomes

$$D_\lambda(x^d, x_i^d) = \prod_{s=1}^{k_d} w_{d,s}(x_s^d, x_{is}^d, \lambda_s) = \prod_{s=1}^{k_d} \lambda_s^{\mathbf{1}(x_{is}^d \neq x_s^d)} \tag{7}$$

with smoothing parameter $\lambda_s \in [0, 1]$ and indicator function $\mathbf{1}(x_{is}^d \neq x_s^d)$, which is equal to one in case $x_{is}^d \neq x_s^d$ and zero otherwise.

[3] More information on the frequency approach and the associated shortcomings can be found in Li and Racine (2007, 188 ff.).

[4] In order to be exact, we assumed that there is no index overlap of s and q as the latter refers to the continuous regressors.

A combination of the product kernels for the continuous and discrete regressors yields the generalized product kernel:

$$K_{\gamma,i} \equiv K_{\gamma,i}(x, x_i) = C_h(x^c, x_i^c) D_\lambda(x^d, x_i^d) \tag{8}$$

where $\gamma = (h, \lambda)$ with $h = (h_1, .., h_{k_c})'$ and $\lambda = (\lambda_1, ..., \lambda_{k_d})'$ using the definitions of (5), (6), and (7).

A discussion of kernel estimation is always accompanied by a discussion on bandwidth selection $\gamma = (h, \lambda)$, as estimation is highly sensitive to the employed bandwidth selection method. In contrast, the choice of the kernel function itself has only minor impacts on the obtained results. There exist many different procedures to choose bandwidths, ranging from rule-of-thumb to cross-validation methods (Li and Racine, 2007, 66 ff.). Fan *et al.* (1995) state that least-squares cross-validation can be trivially adapted from nonparametric regression to local likelihood estimation. Moreover, they emphasize that "plug-in" methods are preferable as they are found to be less variable than cross-validation. In the simulation study and the empirical example, we employ least-squares cross validation due to convenience of implementation.

3 SIMULATION STUDY

We apply our semiparametric NB2 estimator to simulated data and compare its small-sample performance to that of two parametric benchmark models and a nonparametric conditional density estimator (NPCDE) as recently proposed for estimating health care demand by McLeod (2011).[5] In the following, the comparison focuses on model fit in situations where (i) the conditional mean assumption is valid and (ii) the log-linearity assumption fails to hold. The latter might occur in applied research due to complex biological, psychological or economic processes, e.g. due to sharply increasing impacts of a regressor (e.g. number of chronic conditions), stigma or heterogeneous effects with respect to insurance status, age and/or income. In these settings, a researcher might be unable to correctly specify the parametric relationship of the conditional mean of the response and the regressors such that the assumption $\mathrm{E}[y|x] = \exp(x'\beta)$ is invalid.

We simulate data for samples of size $n = 100, 200, 400$ in $R = 100$ repetitions according to the data generating processes in Table 1. We compare the fit of our semiparametric count data model to that of a parametric Poisson (PP) and a parametric negative binomial type 2 (PNB) model with a

[5] In line with McLeod (2011, 1268), the value with the highest predicted probability is taken as the NPCDE outcome prediction, i.e. the conditional mode of the nonparametrically estimated density

log-linear specification of the conditional mean.[6] Certainly, this serves as a simplifying example in order to illustrate the flexibility of the semiparametric count data model against the restrictive parametric specification. In reality, even parametric models that include a set of interaction, quadratic and cubic terms might be inappropriate to model the demand for health care utilization due to heterogeneous effects. For instance, the assumption $x'\beta$ embodied in the conditional mean does neither allow the regression coefficients to change sign nor to vary at all across individuals, cf. Frölich (2006). Moreover, the simulation study demonstrates the ability of the semiparametric estimator to deal with mixed data as two out of the three regressors included are discrete variables.

The model fit refers to out-of-sample predictions as obtained from splitting the data into 50% training and 50% evaluation observations and is assessed by three goodness-of-fit measures: The Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), and the Mean Absolute Error (MAE). The model fit comparison focuses on the models' predictive power with respect to the conditional mean $\mu = \mathrm{E}[y|x]$.

Table 1: Definition of Data Generating Processes

DGP	Distribution	$\log \mu$	σ^{-2}
DGP1	Poisson	$0.3 - 0.5X_{i,c_1} + 1.5X_{i,d_1} - 1X_{i,d_2}$	.
DGP2	Poisson	$0.3 + 0.75X_{i,c_1} + 0.1X_{i,d_1} - 0.1X_{i,d_2}$ $-1.5X^2_{i,c_1} + 2X_{i,c_1}X_{i,d_1} - 1X_{i,c_1}X_{i,d_2}$	.
DGP3	NB2	$1.2 - 0.4X_{i,c_1} + 0.5X_{i,d_1} - 0.8X_{i,d_2}$	7
DGP4	NB2	$0.8 + 2.5X_{i,c_1} + 0.5X_{i,d_1} - 0.1X_{i,d_2}$ $-2.8X^2_{i,c_1} + 0.8X_{i,c_1}X_{i,d_1} + 1.2X_{i,c_1}X_{i,d_2} - 1X^2_{i,d_2}$	7

The mean results on out-of-sample model fit are presented in Tables 2 to 5. In the data settings with correctly specified parametric models (DGP1 and DPG3), the parametric models exhibit the best model fit in terms of all three goodness of fit statistics. This performance is in line with a basic results obtained for maximum likelihood estimation.[7] However, the local likelihood estimator performs relatively well in comparison to the fully parametric alternatives. In DGP1, the largest loss in precision of the local likelihood estimator

[6] Accordingly, the conditional mean function only includes linear terms and omits interaction or polynomial terms.

[7] It can be shown, that the maximum likelihood models have the minimum MSE provided the models are correctly specified (Winkelmann, 2008).

relative to the parametric models is observed for the MSE that amounts to approximately twice the magnitude of the PP's MSE, on average.[8] The results for the MAE and RMSE are more favorable for the semiparametric model since they are on average only 35% to 55% larger than those of the parametric Poisson. In DGP3, the MAE and RMSE of the semiparametric model are on average 43% to 60% larger than those of the PNB whereas the MSE of the LLNB is approximately twice as large as that of the PNB. With increasing sample sizes, the relative loss in precision becomes larger.

In the misspecification scenarios (DGP2 and DGP4), the local likelihood estimator performs best regarding all model fit statistics. In contrast to the parametric estimators, the semiparametric estimator continues to converge if μ is not log-linear in the regressors. Thus, its performance improves relative to the parametric benchmark models with increasing sample size. While the performance of the LLNB and the misspecified PP (in DGP2) is similar for small samples, the local likelihood estimator outperforms the parametric one if the sample size becomes larger. The relative gain in precision ranges between 36% (RMSE) and 56% (MSE) in the largest sample setting. These results can also be confirmed for the negative binomial models in DGP4 where the relative advantage of the local likelihood model over the parametric benchmark is even larger. Again, the MSE is the most optimistic measure for the semiparametric model and ranges between 54% ($n = 100$) and 19% ($n = 400$) of the MSE of the PNB. The most conservative goodness of fit statistic, the MAE, is on average 22% smaller for the LLNB than for the PNB. In the sample with 400 observations, the MAE of the semiparametric negative binomial model becomes even less than half as large as that of the PNB.

A comparison of the oracle fit w.r.t. the precision parameter σ^{-2} shows that the LLNB performs well in comparison to the PNB, even under correct specification of μ (DGP3). Moreover, in DGP4 the LLNB clearly outperforms the parametric model and improves even further with increasing sample size. On average, the MSE for $n = 400$ only amounts to 45% of that of the PNB with similar results observed for the MAE and the RMSE.

As a side note, it can be concluded that the nonparametric density estimator performs poorly w.r.t. out-of-sample predictive power. This result can be confirmed for all goodness-of-fit statistics. In no case, the NPCDE outperforms a parametric or a local likelihood estimator on average, even in the case of functional form misspecification of μ. Moreover, the goodness of fit statistics are substantially larger than those of the semiparametric and parametric models: For instance, the MSE of the NPCDE amounts to five ($n = 100$) or

[8] In line with Frölich (2006), the terms "relative loss" and "relative gain in precision" are used in order to denote the relative difference of the goodness of fit statistics of two estimators

even sixteen ($n = 400$) times the MSE of the PNB. Less extreme results can be confirmed for the MAE, however the MAE is still approximately twice as large as that of the misspecified Poisson model and still exceeds that of the misspecified PNB, on average.

Table 2: Simulation Results, Parametric Poisson

		Model Fit			
		μ			
n	DGP	Bias	MSE	MAE	RMSE
100	DGP1	-0.0266	0.1901	0.2628	0.3945
200	DGP1	0.0177	0.1224	0.2060	0.3180
400	DGP1	0.0002	0.0563	0.1429	0.2189
100	DGP2	-0.0051	0.5432	0.4772	0.7018
200	DGP2	0.0190	0.4049	0.4372	0.6226
400	DGP2	0.0063	0.3414	0.4026	0.5789
100	DGP3	0.0032	0.4014	0.4072	0.5629
200	DGP3	-0.0389	0.1586	0.2698	0.3680
400	DGP3	-0.0126	0.0940	0.2000	0.2796
100	DGP4	0.0539	2.9319	1.1719	1.6709
200	DGP4	0.0152	2.4859	1.0963	1.5544
400	DGP4	0.0331	2.2275	1.0607	1.4835

Table 3: Simulation Results, Parametric NB2

		Model Fit							
		μ				σ^{-2}			
n	DGP	Bias	MSE	MAE	RMSE	Bias	MSE	MAE	RMSE
100	DGP1	- 0.0879	0.2201	0.2905	0.4391	.	.	.	.
200	DGP1	-0.0267	0.0803	0.1805	0.2647	.	.	.	.
400	DGP1	0.0178	0.0536	0.1439	0.2178	.	.	.	.
100	DGP2	-0.0097	0.5712	0.4816	0.7263	.	.	.	.
200	DGP2	0.0195	0.4508	0.4476	0.6543	.	.	.	.
400	DGP2	0.0009	0.3823	0.4122	0.6119	.	.	.	.
100	DGP3	0.0250	0.3971	0.4020	0.5595	0.3348	14.9877	3.3265	3.3265
200	DGP3	-0.0357	0.1617	0.2735	0.3713	1.0032	12.9540	2.8189	2.8189
400	DGP3	-0.0101	0.0924	0.1981	0.2764	0.9081	8.6838	2.2642	2.2642
100	DGP4	0.2340	4.4326	1.2835	2.0036	-2.5575	13.3282	3.3424	3.3424
200	DGP4	0.1611	3.5916	1.1872	1.8553	-3.6002	14.2473	3.6513	3.6513
400	DGP4	0.1950	3.4338	1.1642	1.8208	-3.7532	14.5642	3.7532	3.7532

Table 4: Simulation Results, Local Likelihood NB2

		Model Fit							
		μ				σ^{-2}			
n	DGP	Bias	MSE	MAE	RMSE	Bias	MSE	MAE	RMSE
100	DGP1	-0.0286	0.4189	0.4064	0.5985	.	.	.	.
200	DGP1	0.0132	0.2048	0.2841	0.4279	.	.	.	.
400	DGP1	-0.0064	0.1270	0.2226	0.3331	.	.	.	.
100	DGP2	-0.0106	0.5518	0.4998	0.7034	.	.	.	.
200	DGP2	0.0263	0.2946	0.3714	0.5131	.	.	.	.
400	DGP2	0.0087	0.1505	0.2731	0.3708	.	.	.	.
100	DGP3	-0.0007	0.7549	0.5760	0.7978	-3.0129	15.6420	3.5199	3.5199
200	DGP3	-0.0293	0.3694	0.4191	0.5749	-1.4628	12.1668	3.0638	3.0638
400	DGP3	-0.0219	0.2137	0.3167	0.4382	-0.8660	8.4315	2.5061	2.5061
100	DGP4	0.0558	2.3846	1.0099	1.4573	-3.3451	15.1438	3.5761	3.5761
200	DGP4	0.0209	1.1896	0.7132	1.0464	-2.3671	10.6257	2.8426	2.8426
400	DGP4	0.0562	0.6604	0.5236	0.7797	-1.2981	6.5217	2.1508	2.1508

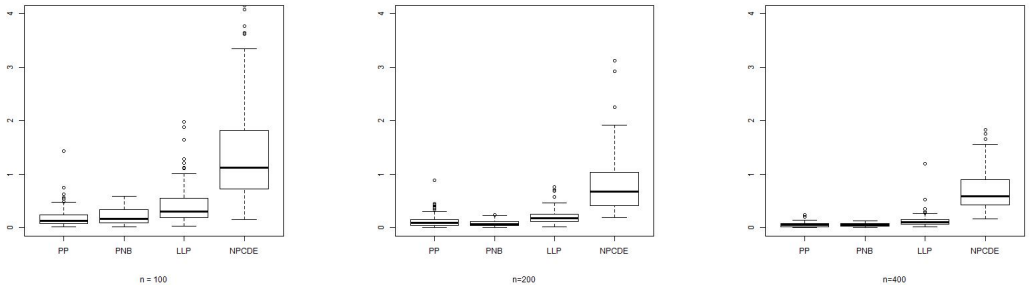
Table 5: Simulation Results, Nonparametric Conditional Density Estimator

		Model Fit			
		μ			
n	DGP	Bias	MSE	MAE	RMSE
100	DGP1	-0.5491	1.5349	0.8260	1.1628
200	DGP1	-0.4296	0.8167	0.6609	0.8586
400	DGP1	-0.4453	0.7007	0.6232	0.8089
100	DGP2	-0.4951	1.6624	0.9097	1.2347
200	DGP2	-0.5398	1.2237	0.8117	1.0714
400	DGP2	-0.5619	0.9402	0.7330	0.9458
100	DGP3	-0.6223	2.0965	1.0858	1.3715
200	DGP3	-0.6977	1.4895	0.9453	1.1763
400	DGP3	-0.8472	1.4995	0.9617	1.1912
100	DGP4	-0.8706	5.9927	1.6018	2.3618
200	DGP4	-0.9621	4.9050	1.4721	2.1588
400	DGP4	-0.9669	3.8918	1.3133	1.9259

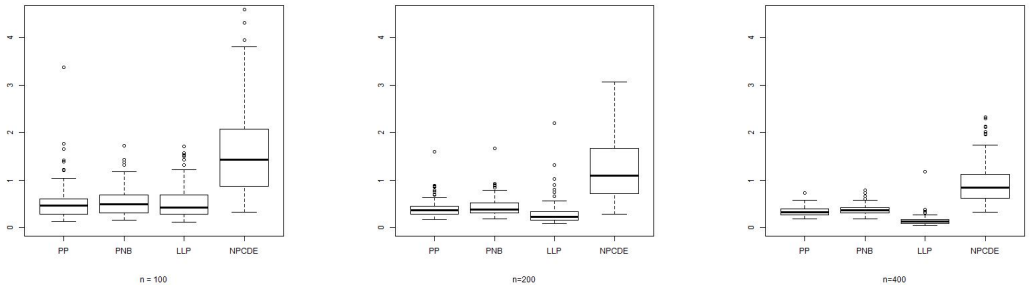
In addition to the average results over the 100 repetitions, we illustrate the results as obtained in every single repetition in the form of boxplots in Figure 1a. This visualization allows to provide further insights on the variability of the MSE of the parametric, semiparametric and nonparametric models. Only if the parametric Poisson and NB2 models are correctly specified (DGP1 and DGP3), the range of their boxplots shrinks with increasing sample size indicating fast convergence in these cases. The average results obtained for the correctly specified parametric estimators shown in Tables 2 to 5 are characterized by a particular degree of robustness. In comparison to the PP and PNB, the local likelihood NB2 model appears to be slightly more variable, although the range of the boxplots decreases with increasing sample size (at a lower rate). In the misspecification cases DGP2 and DGP4, however, the superiority of the LLNB to the other models becomes obvious. The range of its boxplots shrinks not only when the parametric model is correctly specified but also under misspecification. Finally, the boxplots confirm that, in every DGP, the out-of-sample MSE of the NPCDE obtained in the 100 repetitions is much more variable than the goodness-of-fit statistics of the parametric and semiparametric models.

Figure 1: Boxplots, MSE (μ)

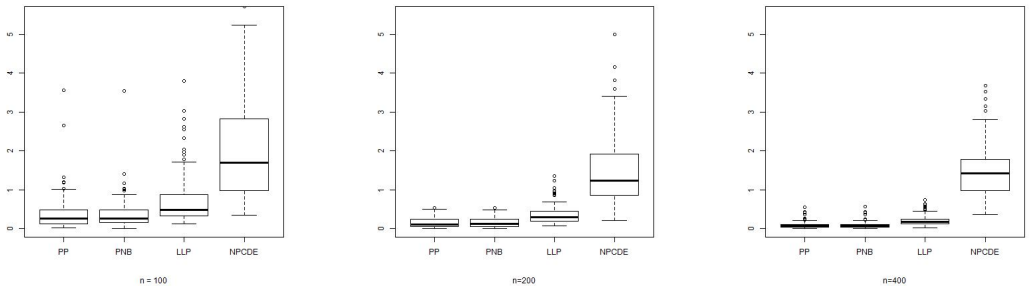
(a) DGP1



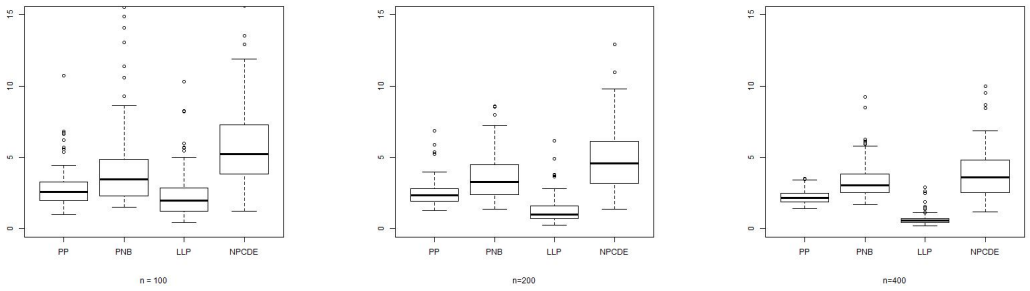
(b) DGP 2



(c) DGP3



(d) DGP4



4 APPLICATION TO HEALTH SERVICE DEMAND

4.1 Data Set and Descriptive Analysis

We apply the local likelihood NB2 estimator to the NMSE data set on health service demand taken from the study by Deb and Trivedi (1997). The model fit of the LLNB estimator is compared to that of frequently employed parametric models and the NPCDE from Section 3. In line with the literature, the focus of the empirical analysis is on in-sample model fit which is assessed by the MSE, MAE and RMSE. The NMES data set is characterized by a considerable variety regarding measures for health service demand and variables of health-related and socio-economic determinants. Accordingly, two different count variables on health care utilization are used for regression analysis.

Table 6: Variable Definitions and Summary Statistics, NMES Data

Statistic	Definition	Mean	St. Dev.
OFP	Number of physician office visits	5.774	6.759
OFNP	Number of nonphysician office visits	1.62	5.32
HLTH	Self-reported health status =0 if health is poor =1 if health is fair =2 if health is excellent	0.281	0.598
NUMCHRON	Number of chronic conditions	1.542	1.350
ADLDIFF	=1 if a person has a condition that limits activities of daily living	0.204	0.403
SITE	Region of residence =1 if northeastern US =2 if midwestern US =3 if western US	1.259	1.134
AGE	Age in years divided by 10	7.402	0.633
BLACK	=1 if a person is African American	0.117	0.322
MALE	=1 if a person is male	0.404	0.491
MARRIED	=1 if a person is married	0.546	0.498
SCHOOL	Number of years of education	10.290	3.739
FAMINC	Family income in $ 10,000	2.527	2.925
EMPLOYED	=1 if the person is employed	0.103	0.304
PRIVINS	=1 if the person is covered by private health insurance	0.776	0.417
MEDICAID	=1 if the person is covered by Medicaid	0.091	0.288

Source: Adapted from Cameron and Trivedi (2013, 230), Table 6.2.

The NMES data from the 1987 and 1988 National Medical Expenditure Survey (NMES) is a subsample of a large representative survey among non-institutionalized individuals and persons that are in long-term care facilities in 1987 (Deb and Trivedi, 1997). It contains information on the health service demand and associated payments of 4,406 U.S. citizens aged 66 and older. The data set bases on quarterly interviews and provides a broad variety on individuals' health service use, health insurance, treatment-related payments,

health status and socio-economic characteristics, cf. Table 6. In the empirical application, we model health care demand for individuals aged 66 and older using two distinct variables on health service utilization: The number of physician office visits, OFP, and the number of nonphysician office visits, OFNP. An attractive feature of the OFP variable is that it is relatively mildly affected by excess zeros as zeros are observed for only 15.5% of the observations. The second variable to assess the health care demand of the elderly, OFNP, covers visits to health care providers who are not physicians, e.g. physiotherapists or respiratory therapists, and has zero-counts for 68% of the sample. The explanatory variables contained in the data set comprise health-related (self-reported health status and limitations in everyday activities), demographic (education, gender, race, age, marital status), economic (family income and employment status) and insurance-related characteristics (private health insurance, insurance provided in the context of medicaid).

4.2 Results

In addition to the local likelihood NB2 (LLNB) estimator, we estimate 13 parametric models and the nonparametric density estimator (NPCDE) as suggested for estimating health care demand by McLeod (2011). The parametric models comprise the parametric Poisson (PP) and four different negative binomial models, i.e. the negative binomial type 1 (NB1) and type 2 (NB2), the NBH and the NBP as suggested by Yee (2014). Moreover, finite mixture versions of the Poisson (FMM Poisson) and the negative binomial 1 (FMM NB1) and 2 (FMM NB2) are implemented.[9] In addition to hurdle extensions of the Poisson (Hurdle Poisson), NB2 (Hurdle NB2) and NBH (Hurdle NBH), a zero-inflated Poisson (Zinf Poisson) and zero-inflated NB2 (Zinf NB2) model are estimated. In all regression models, all explanatory variables listed in Table 6 are included. The count variables OFP and OFNP are used as dependent variables. The bandwidths of the kernel estimators are selected by least squared cross-validation and can be found in the appendix.

In both regression models, the in-sample model fit of all parametric models is highly similar, as the goodness of fit statistics only differ in their decimal places. The results do not change substantially whether the MSE, the MAE or the RMSE is used, both with OFP and OFNP as dependent variable. Although, the Poisson models perform relatively well, these results should be interpreted with caution given the substantial amount of overdispersion in the data. The best performance of a model with negative binomial specification

[9] The predictions for the finite mixture models correspond to the definition in McLeod (2011, 1268), i.e. as the sum of the component predictions weighted by the posterior classification probabilities

Table 7: In-Sample Goodness of Fit Results

Model	OFP			OFNP		
	MSE	MAE	RMSE	MSE	MAE	RMSE
LLNB	36.672[a]	3.973[a]	6.056[a]	25.869[a]	2.229	5.086[a]
PP	41.421	4.167	6.436	27.652	2.285	5.259
NB1	41.619	4.170	6.451	27.720	2.287	5.265
NB2	41.908	4.183	6.474	27.748	2.299	5.268
NBH	41.524	4.167	6.444	27.694	2.289	5.262
NBP	41.632	4.167	6.452	27.710	2.282	5.264
FMM Poisson	41.739	4.229	6.461	27.815	2.366	5.274
FMM NB1	41.437	4.174	6.437	27.707	2.281[b,c]	5.264
FMM NB2	41.997	4.187	6.481	27.727	2.292	5.266
Hurdle Poisson	41.158[b]	4.149[b]	6.415[b]	27.634[b]	2.284	5.257[b]
Hurdle NB2	41.396	4.158	6.434	27.692	2.291	5.262
Hurdle NBH	41.398	4.158	6.434	.*	.*	.*
Zinf Poisson	41.161	4.149	6.416	27.634	2.284	5.257
Zinf NB2	41.298[c]	4.150[c]	6.426[c]	27.650[c]	2.287	5.258[c]
NPCDE	59.386	4.533	7.706	30.876	1.616[a]	5.557

**: Estimator failed to converge*
[a]*: Lowest value of all models*
[b]*: Lowest value of all parametric models*
[c]*: Lowest value of the parametric models with NB specification*

is found for the zero-inflated NB2 with a slightly lower MSE than the hurdle version. The latter, in turn, outperforms the basic NB2 model. The findings of Deb and Trivedi (1997) can be confirmed in that the NB1 model appears to fit the data better than the NB2. This statement holds for both the baseline and the FMM versions.

Overall, the LLNB exhibits the best model fit in the NMES example. This statement holds irrespectively of the outcome variable used and the choice of the goodness-of-fit statistic. As a side result, the application reveals a substantial weakness of the NPCDE in the context of estimating the demand for health care services. While the NPCDE appears highly unattractive in the OFP regression, the performance improves in the OFNP case relative to the other estimators (it even outperforms all other estimators according to the MAE). The variable performance of the NPCDE in two empirically different, although intuitively related, settings can be explained by the method on how predictions are generated.[10] In fact, the NPCDE is effectively a conditional mode estimator and, hence, the model fit is highly sensible to the empirical distribution, i.e. the fraction of zero-observations.

[10] The histograms of the outcome variables can be found in the Appendix as an illustration of the different empirical distribution.

5 CONCLUSION

In the previous sections, a new semiparametric count data model has been proposed in order to model health care demand. The derived local likelihood estimator allows to abstract from the log-linearity assumption imposed on the conditional mean in virtually all common count data models. The semiparametric estimator enables researchers to model heterogeneous effects and allows for consistent estimation even in the case of complex settings in which researchers might fail to correctly specify the parametric conditional mean. Moreover, our semiparametric estimator explicitly addresses (i) overdispersed and (ii) mixed data as currently encountered in estimation of health service demand.

The simulation study and the empirical application to the NMES data set provide encouraging results based on the predictive power of the semiparametric model. The results are characterized by a substantial robustness, whereas the local likelihood estimator is found to benefit from larger data sets. Furthermore, the performance of the semiparametric estimator is shown to be superior to parametric and a nonparametric model even if different goodness-of-fit statistics and measures for health service demand are used. Finally, the sensitivity to the empirical distribution, in particular to the share of zero-observations, is revealed as a weakness of the nonparametric estimator. Although the latter might indeed be a sensible tool in order to detect heterogeneous patterns in the data, it may be of limited use to model health service demand.

Despite the good results of the local likelihood estimators, there is still scope for further improvements. On the one hand, the implemented estimator is a local constant estimator, i.e. it does not benefit from gains regarding bias reduction that are associated with higher-order polynomial approximations (Fan *et al.*, 1995). On the other hand, the bandwidths of the local likelihood estimators have been obtained by least-squares cross-validation for convenience of implementation. For instance, Fan *et al.* (1995) and Frölich (2006) show that the performance of a local likelihood estimator improves if "plug-in" methods of asymptotically optimal smoothing parameters are used.

Future studies might exploit further advantages of the local likelihood approach: Since the likelihood framework provides explicit expressions on the variance of the estimator (Fan *et al.*, 1998), benefits in terms of inference practicability (confidence bounds) might be arguments in favor of the local likelihood estimator in applied research.

6 APPENDIX

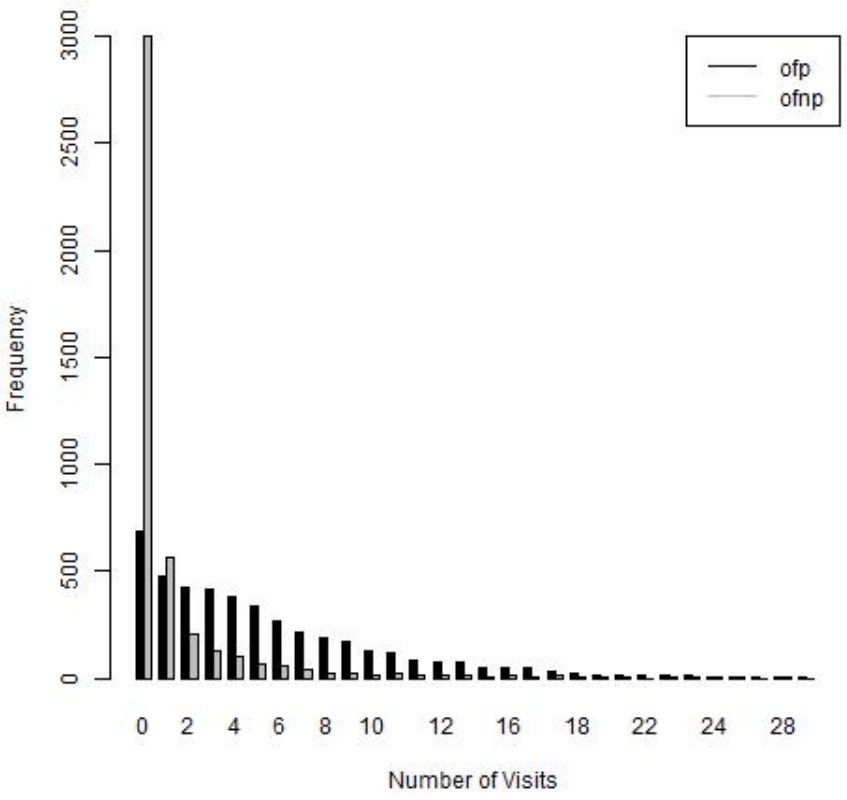
Note on Implementation

The statistical software used in the simulation study is the 3.2.0 version of R in combination with version 0.99.446 of the R project user interface. For estimation of the parametric Poisson model, the `MASS` package, version 7.3-43 by Venables and Ripley (2002). The `VGAM` package, version 1.0-0, by Yee (2015) is used for estimation of the parametric NB2 model.The NPCDE and the local likelihood estimators are implemented by the 0.60-2 version of the `np` package by Hayfield and Racine (2008).

Most estimations of the empirical application are conducted with the 3.2.0 version of R in combination with version 0.99.446 of the R project user interface. For estimation of the parametric Poisson model, the `MASS` package, version 7.3-43 by Venables and Ripley (2002) has been used. The `VGAM` package, version 1.0-0, by Yee (2015) is used for estimation of the parametric NB1, NB2, NBH, NBP and the HNBH model. In order to estimate the hurdle and zero-inflated Poisson and NB2 models, the `pscl` package for R, version 1.4.9., by Jackman (2015) is used. The NPCDE and the local likelihood estimators are implemented by the 0.60-2 version of the `np` package by Hayfield and Racine (2008). Due to the lack of a powerful R module, FMMs are estimated with the `fmm` module by Deb (2012) in STATA 12.0. The previous results obtained for the NMES data set w.r.t. parametric tests and the likelihood function value of Zeileis *et al.* (2008) and Yee (2014) can be replicated. Moreover, the results by Deb and Trivedi (1997) can be replicated using the STATA module `fmm`. Slight deviations occur due to a change in the variable definitions of HLTH and SITE, cf. table 6.

Histogram of the Outcome Variables, NMES Data

Figure 2: Histogram, OFP and OFNP



Note: The histogram shows the absolute frequencies for counts from 0 to 30 for ease of presentation.

Bandwidth Selection: In-Sample Model Fit

Table 8 shows the bandwidths of the LLP and NPCDE in the application using all observations ($n = 4,406$). The bandwidths are obtained by the method of least squares cross validation.

Table 8: Bandwidths for Kernel Estimators, Full Sample

		OFP		OFNP	
Variable	Type	NPCDE	LLP	NPCDE	LLP
OFP / OFNP	ordered	0.000	.	0.000	.
HLTH	ordered	0.784	0.008	0.135	1.000
NUMCHRON	ordered	0.187	0.194	0.482	0.281
ADLDIFF	factor	0.402	0.500	0.500	0.196
SITE	factor	0.676	0.750	0.313	0.349
AGE	continuous	1.672	0.841	3,401,223	1.234
BLACK	factor	0.500	0.500	0.097	0.086
MALE	factor	0.393	0.214	0.187	0.500
MARRIED	factor	0.500	0.378	0.347	0.500
SCHOOL	ordered	0.794	0.777	0.770	0.611
FAMINC	continuous	466,764	21,633,121	26.157	3.112
EMPLOYED	factor	0.250	0.048	0.500	0.087
PRIVINS	factor	0.025	0.136	0.037	0.500
MEDICAID	factor	0.278	0.314	0.323	0.500

In general, the bandwidths chosen are well-behaved, so undersmoothing can be ruled out.[11] The only outstanding bandwidth values computed for the LLNB and NPCDE model are the extremely large smoothing parameters w.r.t. the FAMINC variable in the OFP and OFNP regression models. In addition, the smoothing parameter chosen for the AGE variable in the NPCDE regression of OFNP is very large. Kernel estimators share the property of smoothing-out irrelevant variables (Li and Racine, 2007). Arguing that FAMINC is irrelevant in the regression model of physician office visits is in line with the results of Deb and Trivedi (1997) who find that family income does not affect health service demand in a FMM NB1. In addition, the results of Deb and Trivedi (1997) show that the coefficient of AGE is not significantly different from zero in a FMM NB1.

[11] Undersmoothing refers to a situation with bandwidths that are virtually equal to zero, leading to interpolation of the data points and a perfect model fit (i.e. the bias is equal to zero, but the variance is very large)

7 REFERENCES

Aitchison J, Aitken CG. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* **63**(3): 413–420.

Cameron AC, Trivedi PK. 2013. *Regression Analysis of Count Data*. volume 53 of *Econometric Society Monographs*. Cambridge University Press, Cambridge (U.K.), New York. 2nd edition.

Deb P. 2012. *FMM: Stata module to estimate finite mixture models*.

Deb P, Trivedi PK. 1997. Demand for Medical Care by the Elderly: A Finite Mixture Approach. *Journal of Applied Econometrics* **12**(3): 313–336.

Fan J, Gijbels I. 1996. *Local Polynomial Modelling and Its Applications*. volume 66 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London (U.K.).

Fan J, Heckman NE, Wand MP. 1995. Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. *Journal of the American Statistical Association* **90**(429): 141–150.

Fan J, Farmen M, Gijbels I. 1998. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(3): 591–608.

Frölich M. 2006. Non-parametric regression for binary dependent variables. *The Econometrics Journal* **9**(3): 511–540.

Hayfield T, Racine JS. 2008. *Nonparametric Econometrics: The np Package*.

Hilbe JM. 2011. *Negative Binomial Regression*. Cambridge University Press, Cambridge (U.K.), New York. 2nd edition.

Jackman S. 2015. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. R package version 1.4.9.

Jones AM, Rice N, Bago d'Uva T, Balia S. 2013. *Applied Health Economics*. Routledge, London (U.K.). 2nd edition.

Li Q, Racine JS. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, Princeton, New Jersey.

McLeod L. 2011. A nonparametric vs. latent class model of general practitioner utilization: Evidence from Canada. *Journal of Health Economics* **30**(6): 1261–1279.

Robinson PM. 1988. Root-N-Consistent Semiparametric Regression. *Econometrica* **56**(4): 931–954.

Santos JA, Neves MM. 2008. A local maximum likelihood estimator for Poisson regression. *Metrika* **68**(3): 257–270.

Severini TA, Staniswalis JG. 1994. Quasi-likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association* **89**(426): 501–511.

Staniswalis JG. 1989. The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* **84**(405): 276–283.

Tibshirani R, Hastie T. 1987. Local Likelihood Estimation. *Journal of the American Statistical Association* **82**(398): 559–567.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. Springer, New York. 4th edition.

Weisberg S, Welsh A. 1994. Adapting for the missing link. *The Annals of Statistics* **4**: 1674–1700.

Winkelmann R. 2008. *Econometric Analysis of Count Data*. Springer, Berlin. 5th edition.

Yee TW. 2014. Reduced-rank vector generalized linear models with two linear predictors. *Computational Statistics and Data Analysis* **71**: 889–902.

Yee TW. 2015. *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.0-0.

Zeileis A, Kleiber C, Jackman S. 2008. Regression Models for Count Data in R. *Journal of Statistical Software* **27**(1).