Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

hche | Hamburg Center for Health Economics

Semiparametric Count Data Modeling with an Application to Health Service Demand

Philipp Bach Helmut Farbmacher Martin Spindler

Research Paper
Year: 2017
No: 15

Semiparametric Count Data Modeling with an Application to Health Service Demand

Philipp Bach Helmut Farbmacher Martin Spindler

hche Research Paper No. 15
http://www.hche.de

Abstract

Heterogeneous effects are prevalent in many economic set-tings. As the functional form between outcomes and regressors is generally unknown a priori, a semiparametric negative binomial count data model is proposed which is based on the local likelihood approach and generalized product kernels. The local likelihood framework allows to leave unspecified the functional form of the conditional mean, while still exploiting basic assumptions of count data models (i.e. non-negativity). Since generalized product kernels allow to simultaneously model discrete and continuous regressors, the curse of dimensionality is substantially reduced. Hence, the applicability of the proposed estimator is increased, for instance in estimation of health service demand where data is frequently mixed. An application of the semiparametric estimator to simulated and real-data from the Oregon Health Insurance Experiment provide results on its performance in terms of prediction and estimation of incremental effects.

Keywords: semiparametric regression; nonparametric regression; count data; mixed data; health care demand

JEL classification: C14, C21, I13

Philipp Bach &
Prof. Dr. Martin Spindler

Hamburg Center for Health Economics
Universität Hamburg
Esplanade 36
20354 Hamburg
Germany

info@hche.de

Dr. Helmut Farbmacher

Munich Center for the Economics of Aging
Max Planck Society
Amalienstr. 33
80799 Munich
Germany

farbmacher©mea.mpisoc.mpg.de

1 INTRODUCTION

Estimating the demand for health services is a major field of application of count data regression, since the observed outcome variables of interest only take on non-negative integer values, for instance, the number of visits to a doctor 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' demand for health care. The predominant regression techniques for modeling health service demand are entirely parametric, for example, the Poisson, negative binomial, zero-inflated, and hurdle regression models. Being typically estimated by maximum likelihood, these models incorporate potentially restrictive assumptions which might severely limit the analysis of heterogeneous effects. Basically, the assumptions in parametric count data regression refer to the distribution of the outcome variable or to the parametrization of the conditional mean. While the count data literature has focused on relaxing the distributional assumption, we will concentrate on the conditional mean specification. In virtually all existing count data models used in practice, the assumption $\mathrm{E}[y|x] = \exp(x'\beta)$ is required for consistent estimation. 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). Moreover, the conditional mean assumption substantially limits the analysis of heterogeneous effects, which might be considered in the context of robustness checks or be the major object of interest in an empirical study. For instance, in a policy evaluation, researchers might want to elaborate the direction and magnitude of the impact of a policy measure for some subgroups in the sample in order to prevent potentially countervailing effects and to optimize accordingly the corresponding program.

The conditional mean assumption typically imposed in parametric and semiparametric count data regression specifies a log-linear conditional mean, i.e. $\mathrm{E}[y|x] = \exp(x'\beta)$. In the following, we focus on the linearity imposed on the argument of the exponential function, i.e. the single index $x'\beta$. We still maintain the exponential response function as it guarantees the non-negativity of the dependent variable. In this study, we abstract from a discussion of the exponential function as a valid response function. See Weisberg and Welsh (1994) for a discussion of this topic. A violation of the conditional mean assumption causes inconsistent estimation of almost all count data models. Moreover, the ability to model heterogeneous effects is restricted by the single index specification (Frölich, 2006). As pointed out in Winkelmann (2008), the linearity assumption embodied in the conditional mean might be violated

frequently, for instance, due to nonlinear or heterogeneous effects. In the context of the demand for health services, nonlinearities might arise for certain characteristics. For example, the number of a person's hospital visits might increase more sharply if that person suffers from multiple chronic conditions at the same time, as compared to a situation without any previous disease. Moreover, heterogeneous effects, e.g. heterogeneous responses to a certain policy, might cause a violation of the linearity assumptions and might not be detected by parametric count data models (Winkelmann, 2008; McLeod, 2011). Heterogeneity refers here to different effects for individuals with different characteristics. In the empirical application in Section 4, we provide an example where the effect of providing access to Medicaid differs according to the level of the household income – a pattern that cannot be revealed by using a linear specification of the conditional mean.

In this paper we propose a semiparametric negative binomial type 2 estimator that is based on the local likelihood approach. This allows us to abstract from the linearity assumption embodied in the conditional mean specification and to take into account overdispersion at the same time. The local likelihood approach, as initially developed by Tibshirani and Hastie (1987), is introduced to the context of modeling the demand for health services. 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) but 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 to leave unspecified the relation between the covariates and the conditional mean of the independent variable, and thus allows for potential nonlinearities. For this reason, it is well-suited to uncover heterogeneous effects in the data. Second, it maintains a likelihood structure and, hence, specific estimators for count data regression can be developed. As a consequence, efficiency gains can be achieved compared to fully nonparametric estimators (Frölich, 2006).

This paper contributes to the literature in various ways. It is the first to introduce the local likelihood approach into the field of estimating health care demand. Moreover, it extends the previous work of Santos and Neves (2008) on local likelihood estimation for count data regression, 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 such regression models. Frequently, studies in this field are also interested in a treatment effect, e.g., health insurance provision, with regressors typically defined as binary variables. Furthermore, the local likelihood negative binomial type 2 estimator

derived in this paper is compatible with overdispersed data, i.e. it allows abstracting from the equidispersion assumption maintained in the Poisson model by Santos and Neves (2008). Finally, this paper offers the first goodness-of-fit comparison of a local likelihood estimator for count data regression with commonly implemented fully parametric and nonparametric estimators, in both a simulation study and an empirical application.

Up to now, there have been only a few count data models that allow abstracting from the log-linearity assumption on the conditional mean. Rather, many of these methods focus on the choice of the exponential function as a response function, for instance Weisberg and Welsh (1994). Winkelmann (2008) and Cameron and Trivedi (2013) provide an overview of methods to deal with violations of the conditional mean assumption. In so-called generalized partially linear models (Robinson, 1988), 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 estimating the unknown relation by kernel weighted log-likelihood (Staniswalis, 1989). However, this approach is limited by the necessity of separating the covariates for which a log-linear relation is known from those with an unspecified relation.

Due to the encountered limitations, the existing semiparametric methods may be of limited use in health economic settings in which the validity of the log-linearity assumption is doubted. Alternatively, researchers might employ fully nonparametric methods that do not impose any assumptions on the relation between the dependent variable and the regressors. In a recent study, McLeod (2011) suggests a nonparametric kernel density estimator in order to model health service demand and finds a superior in-sample 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).

As an application, we analyze data from the Oregon Health Experiment (Finkelstein *et al.*, 2012). As a lottery was key in conducting the experiment, for randomizing the possibility of getting health insurance, we are in particular interested in the estimation of the intent-to-treat effect of the result of the lottery. We detect a nonlinear effect and, hence, a heterogeneity in the intent-to-treat effect according to individuals' income. Our results suggest that the effect varies substantially for different levels of income and that it is related to

individuals' eligibility.

The remainder of this paper is organized as follows: In Section 2 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 is *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. sth component of the discrete regressors vector x_i^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 in Winkelmann (2008) (including notation). In contrast to the parametric NB2 framework, the linearity assumption in the specified conditional mean, i.e. $x'\beta$ in $\mathrm{E}[y|x] = \exp(x'\beta)$, is dropped. The conditional probability function of y, $f(y|\mu, \sigma^2)$, is the negative binomial probability function

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

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. $\Gamma(\cdot)$ denotes the Γ-function for which the identity $\Gamma(z + 1) = z\Gamma(z)$ holds which will be employed later in the derivations.

The intuition behind the local likelihood approach can be illustrated best in a comparison of the semiparametric model setup to the parametric framework. In a parametric NB2 model, one would now fully specify the conditional mean as a function of the regressors, $\mu = \exp(x'\beta)$, and maximize it w.r.t. β accordingly. However, in the local likelihood model, we do not assume that the

regressors enter the conditional mean linearly, i.e. $\mathrm{E}[y|x] = \exp(x'\beta)$. Instead, the relation 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 more heavily observations that are close to a certain point (y_i, x_i), a kernel weighting K_γ is introduced in the log-likelihood function. The conditional locally weighted log-likelihood function is set up as

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

In accordance with 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 $\sigma^2(x_0) \approx \sigma_0^2$. Here, for the sake of simplicity of notation, only local constant approximations ($p = 0$) are treated. A more general treatment, with pth order polynomials, can be found in Fan *et al.* (1995) and Fan *et al.* (1998). The $\gamma = (h, \lambda)$ in (2) denotes the vector of smoothing parameters for the continuous (h) and discrete (λ) regressors. The first-order conditions (FOC) w.r.t. μ_0 and σ_0^2 then define the local constant estimators on (μ, σ^2):

$$\frac{\partial \mathcal{L}_0}{\partial \mu_0} = \sum_{i=1}^{n} \left\{ \frac{y_i}{\mu_0} - \frac{(y_i + \sigma_0^{-2})\sigma_0^2}{1 + \sigma_0^2\mu_0} \right\} K_{\gamma,i} = 0 \qquad (3)$$

and

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

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

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} y_i K_{\gamma,i}}{\sum_{i=1}^{n} K_{\gamma,i}}. \qquad (5)$$

Here, the result of Fan *et al.* (1995) on the local likelihood estimation of generalized linear models (GLMs) can be verified, i.e. the expression for the local (constant) likelihood NB2 estimator coincides with that of the Nadaraya-Watson estimator. Accordingly, the local constant likelihood NB2 estimator is

consistent under minimal assumptions, those which are sufficient for the 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 take into account the discrete nature of the regressors. We extend local likelihood estimation so as to smooth discrete variables. This greatly extends its applicability, since discrete variables are often encountered in models of health service demand (for instance, insurance status or treatment evaluations in general). Building upon Li and Racine (2007), the so-called *generalized product kernels* will be discussed in the following. The main advantage of using these kernel functions is that we can use all observations in semi- and nonparametric estimation, instead of fitting the data separately for all possible combinations of the discrete regressors. Therefore the curse of dimensionality only includes the continuous variables and, thus, is substantially less severe than in early versions of kernel regression where the so-called "frequency approach" was used. More information on the frequency approach and its shortcomings can be found in Li and Racine (2007, 188 ff.). 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 by

$$C_h(x^c, x_i^c) = \prod_{q=1}^{k_c} h_q^{-1} w_c \left(\frac{x_q^c - x_{iq}^c}{h_q} \right), \tag{6}$$

where $h_q \in (0, \infty)$ is the bandwidth or smoothing parameter for the 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 the standard assumptions listed in Li and Racine (2007, 9). In the Monte Carlo simulation and the application in Section 3 and 4 we use a Gaussian kernel.

For the discrete regressors x_s^d, with $s = 1, ..., k_d$, we define a product kernel with smoothing parameter $\lambda_s \in [0, 1]$ that incorporates a variation of the kernel function of Aitchison and Aitken (1976), such that

$$w_{d,s}(x_s^d, x_{is}^d, \lambda_s) = \begin{cases} 1, & \text{if } x_{is}^d = x_s^d \\ \lambda_s, & \text{otherwise} \end{cases} . \tag{7}$$

Accordingly, the product kernel for the discrete regressors becomes

$$D_\lambda(x^d, x_i^d) = \prod_{s=1}^{k_d} w_{d,s}(x_s^d, x_{is}^d, \lambda_s) = \prod_{s=1}^{k_d} \lambda_s^{\mathbf{1}(x_{is}^d \neq x_s^d)}, \tag{8}$$

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 when $x_{is}^d \neq x_s^d$ and zero otherwise. A combination of the product kernels for the continuous and discrete regressors yields the generalized product kernel:

$$K_{\gamma,i} \equiv K_{\gamma,i}(x, x_i) = C_h(x^c, x_i^c) D_\lambda(x^d, x_i^d), \tag{9}$$

where $\gamma = (h, \lambda)$ with $h = (h_1, .., h_{k_c})'$ and $\lambda = (\lambda_1, ..., \lambda_{k_d})'$ using the definitions of (6), (7), and (8).

A discussion of kernel estimation is always accompanied by a discussion on the selection of the bandwidth $\gamma = (h, \lambda)$, since estimation is highly sensitive to the employed bandwidth selection method. In contrast, the choice of the kernel function itself has only a minor effect on the results. There are many different procedures for choosing the bandwidths, ranging from rule-of-thumb, to cross-validation (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 in Sections 3 and 4, we employ least-squares cross validation due to its convenience of implementation. More information about the implementation in R can be found in the Appendix and the replications files are available.

3 SIMULATION STUDY

In this section we apply the semiparametric NB2 estimator to simulated data and compare its small-sample performance to that of the parametric benchmark model and a nonparametric conditional density estimator (NPCDE) as recently proposed for estimating health care demand by McLeod (2011). The NPCDE is implemented as suggested in McLeod (2011, 1268), i.e. the value with the highest predicted probability, which corresponds to the conditional

mode of the nonparametrically estimated density, is taken as the NPCDE outcome prediction. In the following, the comparison focuses on estimation of incremental effects and out-of-sample predictive accuracy of the conditional mean. We generate situations where the linearity assumption in $\exp(x'\beta)$ holds and fails to hold. The data generating processes are presented in Table 1. The regressors X_{i,c_1}, X_{i,d_1}, and X_{i,d_2} are drawn from identical distributions across all DGPs. The continuous regressor X_{i,c_1} is drawn from a uniform distribution on the interval $[0, 1]$. The dummy X_{i,d_1} assumes the values 0 and 1 with probability $p_0 = p_1 = 0.5$. X_{i,d_2} takes the values 0, 1 and 2 with equal probability $p_0 = p_1 = p_2 = 1/3$ and is treated as a categorical variable.

Table 1: Definition of the Data Generating Processes

DGP	Distribution	μ	σ^{-2}	p_0
DGP1	NB2	$\exp\left(1.2 - 0.4X_{i,c_1} + 0.5X_{i,d_1} - 0.8X_{i,d_2}\right)$	7	–
DGP2	NB2	$\exp\big(0.8 + 2.5X_{i,c_1} + 0.5X_{i,d_1} - 0.1X_{i,d_2}$ $-2.8X^2_{i,c_1} + 0.8X_{i,c_1}X_{i,d_1} + 1.2X_{i,c_1}X_{i,d_2} - 1X^2_{i,d_2}\big)$	7	–
DGP3	ZiNB2	$(1 - p_0) \cdot \mu_{DGP1}$	7	0.2
DGP4	ZiNB2	$(1 - p_0) \cdot \mu_{DGP2}$	7	0.2

Throughout the simulation study, we implement six different estimators. An overview on the estimated models is given in Table 11 in the Appendix. Additional to the semiparametric local likelihood negative binomial estimator (LLNB) and the NPCDE, we esimate the parametric negative binomial 2 model in a linear conditional mean specification (PNB (1)), i.e. the independent variables enter the regression model only via linear terms. The model PNB (1) is correctly specified under DGP1 and misspecified under DGP2 as all interaction terms are omitted. Additionally, we estimate the parametric NB2 in a more flexible specification (PNB (2)), i.e. we include all two-way interactions of the variables plus a quadratic term of the continuous regressor X_{i,c_1}. This model specification is implemented in order to present a comparison of the performance of a flexible – although still not entirely correctly specified – parametric model (i.e. the model does not include X^2_{i,d_2} in DGP2) with the semiparametric model. In order to assess the performance of the local likelihood estimator in presence of excess zeros, we provide additional results for zero-inflated versions of DGP1 and DGP2, named DGP3 and DGP4. Accordingly, an additional zero-generating process is introduced which sets outcome variable Y_i equal to zero with probability $p_0 = 0.2$. If Y_i is not set equal to zero in this stage, it is generated by a negative binomial distribution as in DGP1 or DGP2. DGP3 and DGP4 generate on average 42% to 45% of zeros

as compared to 27% to 31% of zero-outcomes in DGP1 and DGP2. The models PZNB (1) and PZNB (2) are zero-inflated versions of PNB (1) and PNB (2) with PZNB (1) being correctly specified in DGP 3.

Prediction of Conditional Mean
The first part of the simulation study focuses on the models' predictive power with respect to the conditional mean μ =E$[y|x]$. We generate small samples of size $n = 100, 200, 400$ in $R = 500$ repetitions according to DGPs 1 to 4. Model fit refers to out-of-sample predictions obtained from a 50% data split and is assessed by the mean squared error (MSE), the root mean squared error (RMSE), and the mean absolute error (MAE). Results as averaged over all repetitions are presented in Tables 2 to 7. In the setting with the correctly specified parametric model (DGP1), the PNB (1) exhibits the best model fit in terms of all three goodness of fit statistics. This performance is in line with a basic result obtained for maximum likelihood estimation. It can be shown that the maximum likelihood models have the minimum MSE provided the models are correctly specified (Winkelmann, 2008). However, the local likelihood estimator performs relatively well in comparison to the fully parametric alternative with a MAE and RMSE being 48% to 64% larger than those of the PNB (1) on average.

Table 2: Simulation Results, Parametric NB2, PNB (1)

		Model Fit							
		μ				σ^{-2}			
n	DGP	Bias	MSE	MAE	RMSE	Bias	MSE	MAE	RMSE
100	DGP1	0.0141	0.3614	0.3903	0.5392	0.2956	11.5130	2.8063	2.8063
200	DGP1	0.0036	0.1466	0.2562	0.3518	0.8774	10.9737	2.6863	2.6863
400	DGP1	0.0068	0.0740	0.1835	0.2535	0.7238	8.6618	2.2535	2.2535
100	DGP2	0.2454	4.6975	1.3525	2.0381	-2.7162	13.1709	3.3408	3.3408
200	DGP2	0.1784	3.8201	1.2089	1.8780	-3.3596	13.8310	3.4948	3.4948
400	DGP2	0.1725	3.3418	1.1557	1.8000	-3.7384	14.6295	3.7384	3.7384
100	DGP3	0.0367	1.3282	0.7497	1.1198	-3.7689	20.8667	4.2755	4.2755
200	DGP3	-0.0023	1.1136	0.6877	1.0441	-4.5797	23.4647	4.7010	4.7010
400	DGP3	-0.0064	1.0220	0.6542	1.0055	-4.9672	25.1539	4.9696	4.9696
100	DGP4	0.3339	7.4476	1.6793	2.6173	-5.4385	31.0786	5.5053	5.5053
200	DGP4	0.2299	6.2587	1.5254	2.4429	-5.7779	33.5959	5.7779	5.7779
400	DGP4	0.1994	5.7027	1.4686	2.3642	-5.8897	34.7495	5.8897	5.8897

In the misspecification scenario (DGP2), the local likelihood estimator performs best in terms of all model fit statistics even in comparison to the more flexible parametric negative binomial model PNB (2). While the flexible specification of the parametric pays off in terms of a better out-of-sample predictive performance compared to the linearly specified PNB (1), the MSE, MAE and RMSE of the PNB (2) are substantially larger than those of the LLNB.

Table 3: Simulation Results, Parametric NB2, PNB (2)

		Model Fit							
		μ				σ^{-2}			
n	DGP	Bias	MSE	MAE	RMSE	Bias	MSE	MAE	RMSE
100	DGP1	0.0337	0.8824	0.5678	0.8286	1.0087	12.8673	2.9082	2.9082
200	DGP1	0.0146	0.3149	0.3647	0.5259	1.5038	13.3566	2.9154	2.9154
400	DGP1	0.0082	0.1512	0.2574	0.3702	1.1494	10.0914	2.4265	2.4265
100	DGP2	0.2548	4.0460	1.3438	1.8839	-0.6497	9.9417	2.6363	2.6363
200	DGP2	0.1576	2.8609	1.1010	1.6019	-1.7387	8.6040	2.6043	2.6043
400	DGP2	0.1230	2.0538	1.0146	1.4074	-2.5135	8.3882	2.6944	2.6944
100	DGP3	0.0817	2.0499	0.9058	1.3533	-3.0672	18.9067	4.0236	4.0236
200	DGP3	0.0022	1.3040	0.7537	1.1293	-4.2267	21.1753	4.4074	4.4074
400	DGP3	-0.0030	1.1062	0.6841	1.0456	-4.8248	23.9304	4.8336	4.8336
100	DGP4	0.4435	7.5325	1.8181	2.6294	-4.8439	26.5794	5.0085	5.0085
200	DGP4	0.2098	5.6230	1.4572	2.3142	-5.4983	30.7742	5.5106	5.5106
400	DGP4	0.1646	4.9351	1.3360	2.1924	-5.7249	32.8694	5.7249	5.7249

Table 4: Simulation Results, Parametric Zero-Inflated NB2, PZNB (1)

		Model Fit							
		μ				σ^{-2}			
n	DGP	Bias	MSE	MAE	RMSE	Bias	MSE	MAE	RMSE
100	DGP3	0.0649	1.2803	0.7326	1.0960	0.6909	17.8513	3.4707	3.4707
200	DGP3	-0.0112	1.1047	0.6838	1.0399	1.2994	14.9476	3.0492	3.0492
400	DGP3	0.0098	1.0271	0.6489	1.0083	0.7202	9.7396	2.4797	2.4797
100	DGP4	0.1896	6.3930	1.6121	2.4482	-2.3087	15.1158	3.5010	3.5010
200	DGP4	0.0911	5.4254	1.4833	2.2843	-3.0840	14.6203	3.5238	3.5238
400	DGP4	0.0778	5.0016	1.4390	2.2188	-3.7358	15.7922	3.7948	3.7948

Table 5: Simulation Results, Parametric Zero-Inflated NB2, PZNB (2)

		Model Fit							
		μ				σ^{-2}			
n	DGP	Bias	MSE	MAE	RMSE	Bias	MSE	MAE	RMSE
100	DGP3	0.1375	1.8671	0.8696	1.2891	1.1098	15.9169	3.2548	3.2548
200	DGP3	0.0027	1.2394	0.7333	1.1011	1.1667	12.6048	2.9188	2.9188
400	DGP3	0.0114	1.0872	0.6730	1.0374	1.6609	14.8457	3.0775	3.0775
100	DGP4	0.5287	5.9245	1.8776	2.3705	-0.4685	11.5516	2.7957	2.7957
200	DGP4	0.0596	4.6952	1.3948	2.1182	-0.9419	10.7627	2.7499	2.7499
400	DGP4	0.0192	4.1172	1.2967	2.0121	-2.0107	8.8397	2.6257	2.6257

Table 6: Simulation Results, Local Likelihood NB2, LLNB

		Model Fit							
		μ				σ^{-2}			
n	DGP	Bias	MSE	MAE	RMSE	Bias	MSE	MAE	RMSE
100	DGP1	-0.0130	0.7175	0.5764	0.7762	-2.6566	16.5093	3.6493	3.6493
200	DGP1	-0.0002	0.3727	0.4163	0.5654	-1.4546	11.3117	2.8724	2.8724
400	DGP1	0.0035	0.1929	0.3007	0.4139	-0.7426	8.2746	2.3744	2.3744
100	DGP2	0.0162	2.5268	1.0270	1.5093	-3.0952	15.9351	3.6791	3.6791
200	DGP2	0.0258	1.2522	0.7247	1.0654	-2.0237	10.2281	2.8215	2.8215
400	DGP2	0.0448	0.6344	0.5199	0.7699	-1.2310	6.3926	2.1132	2.1132
100	DGP3	0.0254	1.6898	0.8838	1.2633	-5.4854	32.1597	5.5965	5.5965
200	DGP3	-0.0031	1.3283	0.7741	1.1371	-5.4626	30.4489	5.4764	5.4764
400	DGP3	-0.0081	1.1471	0.7093	1.0642	-5.3692	29.0570	5.3692	5.3692
100	DGP4	0.0520	5.6778	1.5411	2.3252	-5.9819	36.2145	5.9831	5.9831
200	DGP4	0.0217	4.3151	1.3201	2.0528	-5.8178	34.0891	5.8178	5.8178
400	DGP4	0.0228	3.6870	1.1783	1.9062	-5.6631	32.2180	5.6631	5.6631

Table 7: Simulation Results, Nonparametric Conditional Density Estimator, NPCDE

		Model Fit			
		μ			
n	DGP	Bias	MSE	MAE	RMSE
100	DGP1	-0.6424	2.1395	1.0977	1.3962
200	DGP1	-0.6919	1.5581	0.9643	1.2010
400	DGP1	-0.7474	1.3180	0.9042	1.1162
100	DGP2	-0.9036	6.2993	1.6571	2.4230
200	DGP2	-1.0005	5.0307	1.4819	2.1806
400	DGP2	-0.9426	3.9165	1.3189	1.9315
100	DGP3	-1.0113	3.8093	1.4157	1.9154
200	DGP3	-1.3167	4.0350	1.4626	1.9885
400	DGP3	-1.4749	4.4067	1.5371	2.0902
100	DGP4	-1.7878	12.0401	2.2362	3.4141
200	DGP4	-1.9885	12.4984	2.2623	3.5048
400	DGP4	-2.2346	13.4943	2.3707	3.6599

On average, the MSE achieved by the semiparametric estimator amounts up to only 31% of the parametric model's MSE ($n = 400$). As is the case for the comparison with the linearly specified PNB (1), the relative performance gains achieved by the semiparametric model tend to increase with larger sample sizes. A comparison of the model fit in terms of the precision parameter σ^{-2} shows that the LLNB performs well in comparison to the PNB (1) and PNB (2), even under correct specification of μ (DGP1). The LLNB estimates on σ^{-2} benefits from larger samples. In DGP2, the LLNB outperforms both parametric models with sample size $n = 400$.

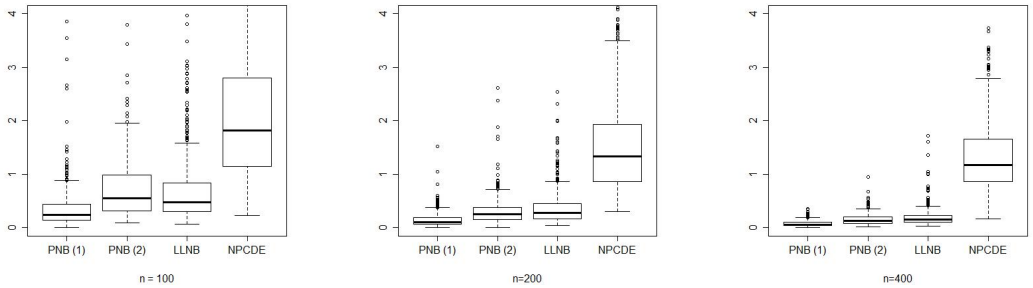
In the settings with excess zeros, the results obtained for the LLNB are still encouraging with an RMSE and MAE being on average ca. 6% to 21% larger than the PZNB (1)'s RMSE and MAE in DGP3. The performance of the LLNB appears to be relatively robust to the existence of excess zeros. Under misspecification of μ (DGP4), the local likelihood estimator outperforms the zero-inflated parametric estimators in terms of MSE, RMSE and MAE irrespective of the sample size. It can be concluded that the presence of excess zeros does not *per se* confound the performance of the LLNB. An analogous conclusion cannot be drawn for the NPCDE that severely suffers from excess zeros in terms of predictive performance. In terms of practicability, it is worth to notice a point related to implementation. The flexibly specified parametric zero-inflated NB2 estimator suffers from convergence problems, in particular if the sample size is small, leading to 304 convergence failures in 500 repetitions for $n = 100$ observations under DGP4.

As a side note, it can be concluded that the nonparametric density estimator performs poorly in terms of out-of-sample predictive power, as basically all goodness-of-fit statistics are by far larger than those of the other models. In no case does the NPCDE outperform a parametric or a local likelihood estimator on average, even in the case of misspecification of the functional form of μ. Moreover, the NPCDE appears to be highly sensitive to excess zeros leading to a particularly bad performance in terms of out-of-sample predictions.

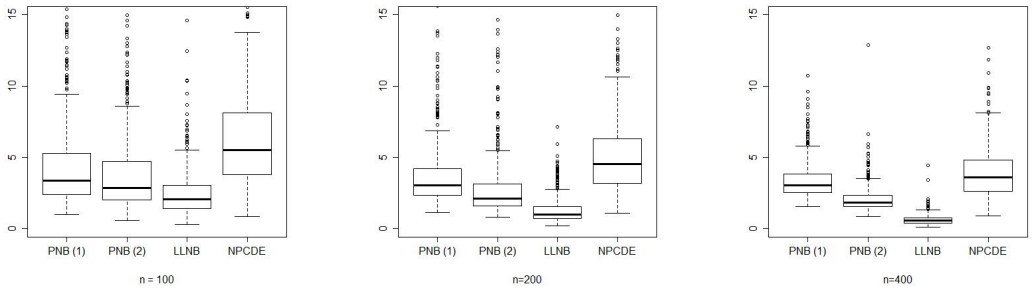
Additionally, we present boxplots of the MSEs computed in every repetition in Figures 1a to 2b to illustrate the robustness of the simulation results. Compared to the PNB (1), the LLNB appears to be slightly more variable in DGP1. However, it can be observed that the LLNB exhibits an almost identical behavior in all DGPs. This behavior cannot be confirmed for the PNB (1), which has a far more variable MSE in DGP2. It becomes obvious that, in contrast to the parametric NB2, the local likelihood model does not require an ex ante specification of the functional form of the conditional mean. While the MSE of the parametric NB2 model (PNB (1) and PNB (2)) is found to be highly variable in the misspecification scenario (DGP2), the semiparametric model continues to converge. Overall, inspection of the boxplots suggests that

Figure 1: Boxplots, MSE (μ), DGP1 and DGP2

(a) DGP1



(b) DGP2



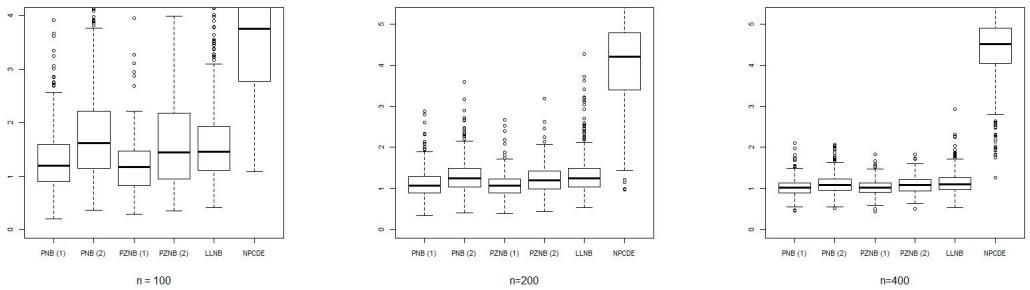
the simulation results are characterized by a particular degree of robustness, even in the presence of a large fraction of zero-counts. Finally, the boxplots confirm that, irrespective of the DGP, the out-of-sample MSE of the NPCDE obtained in the 500 repetitions is much more variable than the goodness-of-fit statistics of the parametric and semiparametric models. This is particularly true in the presence of excess zeros.

Estimation of Incremental Effects

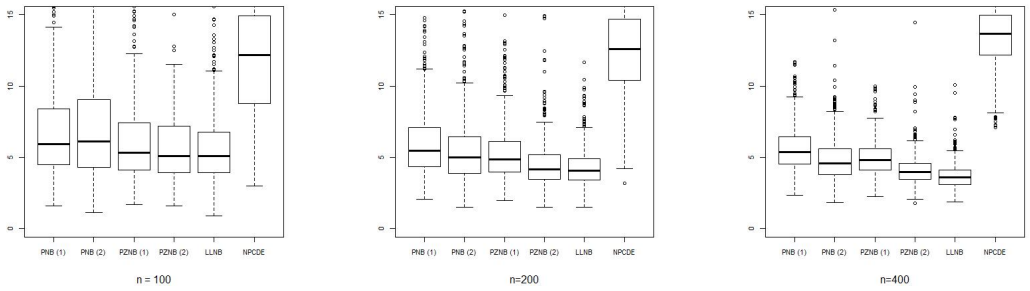
Since researchers are often interested in estimating the impact of some policy program or some treatment effect, the second part of the simulation study concentrates on estimating the incremental effect of the dummy variable X_{i,d_1} from a sample of size $n = 400$ generated by DGP1 and DGP2, respectively. The incremental effect (IE) of variable X_{i,d_1} is defined by $IE(k, l|x) = \mathbf{E}[Y|X = x, X_{i,d_1} = k] - \mathbf{E}[Y|X = x, X_{i,d_1} = l]$, where the other regressors X are fixed at some representative value. In our simulation example, the levels of the treatment dummy X_{i,d_1} are (naturally) $k = 1$, $l = 0$. The true incremental effects is fixed for a representative observation with covariate $X_{i,d_2} = 0$ and

Figure 2: Boxplots, MSE (μ), DGP3 and DGP4

(a) DGP3



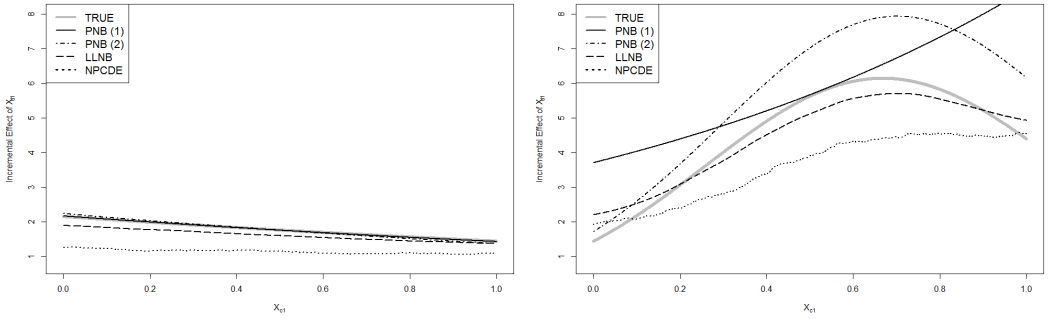
(b) DGP4



X_{i,c_1} held constant at a particular point on a grid from 0 to 1 by step size 0.001 (X_{i,c_1} is drawn from a uniform distribution from 0 to 1). Figures 3a and 3b and Table 8 display the average results obtained from estimation of the incremental effect from every implemented model under DGP1 and DGP2. The plots show the incremental effects at given values of X_{i,c_1} as averaged over all 500 repetitions. In DGP1, the IEs estimated by the parametric models are very close to the true effects (grey line). However, in the case of a nonlinear incremental effect, the linearly specified parametric model, PNB (1), entirely misses the underlying patterns, i.e. the heterogeneity of the true incremental effect w.r.t. X_{i,c_1} (grey line). Despite the ability of the more flexibly specified parametric model, PNB (2), to uncover the nonlinearity of the IE, the resulting estimates are relatively far off the true line. The incremental effect line of the semiparametric model is closest to the true curve in DGP2 which is confirmed by results on the intergrated mean squared and absolute error in Table 8. The results suggests that although a parametric model might be used to detect heterogeneous effects, the quality of the resulting estimates might be severely limited and conclusions might be drawn with caution.

Figure 3: Incremental Effects

(a) DGP1 (b) DGP2



Figures 3a and 3b present the average incremental effect of dummy X_{i,d_1} as estimated by the implemented models on the basis of samples with $n = 400$ observations in $R = 500$ repetitions. In each repetition, incremental effects are estimated at all grid points of X_{i,c_1}. The estimates at a particular grid point are then averaged over all repetitions. The underlying grid is constructed by steps from 0 to 1 of size 0.001. The grey lines present the true incremental effects.

Table 8: Average Results for IMSE and IMAE for Incremental Effects

	PNB (1)		PNB (2)		LLNB		NPCDE	
DGP	*IMSE*	*IMAE*	*IMSE*	*IMAE*	*IMSE*	*IMAE*	*IMSE*	*IMAE*
DGP1	0.1125	0.2655	0.2788	0.4091	0.3494	0.4597	2.2379	1.2704
DGP2	3.3074	1.3961	2.8711	1.3668	1.1382	0.8345	7.2624	2.1934

Table 8 shows the average results of the integrated mean squared error (IMSE) and integrated mean absolute error (IMAE), both calculated on a grid of $X_{i,c_1} = x_c,\ \ x_c \in [0, 1]$ with steps of size 0.001. In every repetition we compute the integrated mean squared error as a measure of distance between the estimated and the true regression lines over all (grid) points of the continuous covariate X_{i,c_1}.

4 APPLICATION TO HEALTH SERVICE DEMAND

4.1 Dataset and descriptive statistics

We use data from the Oregon health insurance experiment, a large scale experiment providing randomly assigned access to public health insurance (Finkelstein *et al.*, 2012). In 2008, more than 85,000 persons signed up to a waiting list for Medicaid in the state of Oregon, USA. Out of this group, approximately 30,000 households were randomly assigned to the treatment group. Treatment status refers to access to Medicaid, i.e. the "winners" of the lottery were given the opportunity to apply for Medicaid. An equal number of individuals were chosen from the waiting list to form the control group. In their extensive study, Finkelstein *et al.* (2012) show that being treated increased an individual's probability to have health insurance by approximately 25 percentage points.

The insurance program which was the subject of the lottery was the Oregon Health Plan (OHP) Standard, a public health program providing relatively generous benefits to adult persons with low income who were not categorically eligible for public insurance (at the time of the experiment, a second public insurance program existed, namely, OHP Plus providing health insurance for certain population groups, e.g. disabled persons or pregnant women). The eligibility criteria of OHP Standard consisted of age (19 to 64 years), residence in Oregon (USA), citizenship or status as a legal immigrant, absence of health insurance for at least six months, and income below the federal poverty level (FPL). Moreover, the individual's assets must not have exceeded $\$2,000$. Being covered by OHP Standard, individuals gained access to comparatively generous benefits, with no consumer cost sharing. The benefits covered most medical procedures, e.g., physician services, prescription drugs, and hospital stays. Dental care and vision care were not covered. For more details on the Oregon lottery, the health insurance programs, the randomization procedure, and the composition of the samples, refer to Finkelstein *et al.* (2012), Baicker *et al.* (2013), Allen *et al.* (2010) and Taubman *et al.* (2014).

Following Finkelstein *et al.* (2012), we estimate the intent-to-treat (ITT) effect of winning the lottery using data from the 12-months follow-up survey. As we extend the set of regressors included in the regression model, we need to discard observations with missing information. As a consequence, our sample size is reduced to $15,518$ complete observations compared to $23,441$ observations in Finkelstein *et al.* (2012). Table 9 compares the demographic characteristics for our subsample with those of the original sample. Overall, the means of the observed demographic characteristics (such as sex, age, income, education) and the number of chronic conditions are very much the same in both samples. We created the number of chronic conditions out of the survey items "have you ever been told you have ... a) diabetes b) asthma c) high blood pressure d) chronic obstructive pulmonary disease e) depression f) heart disease g) congestive heart failure h) high cholesterol or i) kidney problems." Health care utilization, however, seems to be slightly lower in our subsample.

4.2 Results

Our goal is to model the demand for health care by estimating the intent-to-treat effect for the dataset described in the previous section. We use parametric and semiparametric negative binomial regressions to estimate health service demand measured by the number of doctor visits in the last six month (y_{ih}). We follow Finkelstein *et al.* (2012, 1071) where the demand for health care in a linear regression model is given by:

$$y_{ih} = \beta_0 + \beta_1 LOTTERY_h + X_{ih}\beta_2 + \epsilon_{ih}, \tag{10}$$

Table 9: Means of Demographic Characteristics

	subsample	full sample
% Female	0.603	0.592
% White	0.847	0.824
% Black	0.030	0.034
% Spanish/Hispanic/Latino	0.096	0.116
% English preferred language	0.932	0.917
% MSA	0.744	0.748
Age		
% 20–50	0.688	0.669
% 50–64	0.312	0.331
Income (% federal poverty limit)		
$< 50\%$	0.395	0.404
$50\% - 75\%$	0.126	0.129
$75\% - 100\%$	0.147	0.145
$100\% - 150\%$	0.192	0.186
$> 150\%$	0.139	0.136
Education		
% Less than high school	0.146	0.168
% High school diploma or GED	0.502	0.498
% Vocational training or 2-year degree	0.231	0.221
% 4-year college degree or more	0.121	0.113
Insurance coverage		
Any insurance?	0.410	0.411
OHP/Medicaid	0.213	0.215
Private insurance	0.028	0.026
Health status		
Number of chronic conditions	1.410	1.405
	(1.449)	(1.453)
Health Care Utilization		
Outpatient visits last six months	1.815	1.949
	(2.655)	(2.923)
Emergency room visits last six months	0.390	0.439
	(0.903)	(0.969)
Inpatient hospital admissions last six months	0.081	0.096
	(0.365)	(0.399)
Prescription drugs currently	2.144	2.330
	(2.748)	(2.850)
Maximum number of observations	15,518	23,441

Full sample refers to Table V in Finkelstein *et al.* (2012).
Standard deviations in parentheses.

with indices i, h referring to individual i and household h. Whereas the set of regressors, X_{ih}, in Finkelstein *et al.* (2012) is restricted to those covariates being correlated with the probability of winning the lottery (e.g. dummies on household size and survey wave as well as their interactions), we consider additional regressors. We include variables on gender, household income (as a

percentage of the federal poverty level), educational attainment, race (white, black, hispanic) and dummies indicating metropolitan statistical areas and English as preferred survey language. In the linear setup, β_1 can be interpreted as the effect of extending public health insurance coverage on the corresponding outcome variable. We define the ITT effect as the incremental change contrasting the individual counterfactual predictions of winning the lottery with losing:

$$\begin{aligned} ITT(x_{ih}) &= \mathbf{E}[y_{ih}|LOTTERY_h = 1, X_{ih} = x_{ih}] \\ &\quad - \mathbf{E}[y_{ih}|LOTTERY_h = 0, X_{ih} = x_{ih}], \end{aligned}$$

depending on the other regressors X_{ih}. The variables that are included in addition to the regressors in Finkelstein *et al.* (2012) might be correlated with the lottery status and health service demand, so that the ITT effect might differ across individuals. The estimates of β_1 using Finkelstein *et al.* (2012)'s set of regressors and a linear model are 0.269 (0.045) in our subsample and 0.314 (0.054) in Finkelstein *et al.* (2012)'s sample. Standard errors are in parentheses. The full results are available upon request.

Table 10: Estimates from Parametric NB2 Regression

	(1)	(2)	(3)	(4)
Lottery (ITT)	0.147***	0.140***	0.262***	0.185
	(0.023)	(0.022)	(0.033)	(0.158)
Income (% federal poverty line)		-0.000**	0.000	0.000
		(0.000)	(0.000)	(0.001)
Income squared				0.000
				(0.000)
ITT × Income (% federal poverty line)			-0.002***	-0.004***
			(0.000)	(0.001)
ITT × Income squared				0.000**
				(0.000)
Female		0.305***	0.303***	0.293***
		(0.023)	(0.023)	(0.032)
Age		0.009***	0.009***	0.006***
		(0.001)	(0.001)	(0.001)
Log-Likelihood	-28,243.28	-28,065.66	-28,052.68	-28,027.59

Number of observations in all regressions is 15,518. Standard errors in parentheses. *p<0.05; **p<0.01; ***p<0.001. All regressions include the treatment dummy as well as variables on the survey waves, household size and their interactions. Additional regressors included in Regressions (2)-(4) are income, age, and dummies for female, high school or GED, vocational training/2-yr degree, 4-yr degree or more, English preferred language, a dummy on metropolitan statistical area, white, black, and Hispanic. In addition to the regressors reported in Column (4), the interactions of the treatment dummy with all (non-technical) covariates are included in Regression (4), i.e. with gender, income, age, education, English, metropolitan statistical area, race variables and income squared.

We contrast the ITT effect as estimated by semiparametric local likelihood

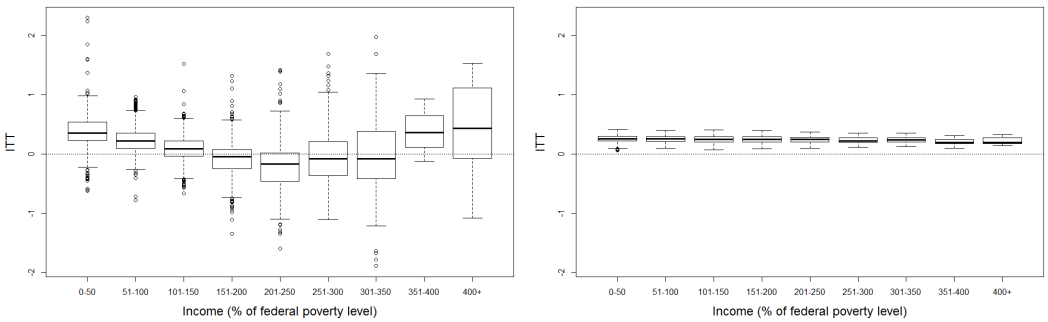
negative binomial 2 estimator (LLNB) to the results obtained from the parametric negative binomial model in various specifications. Table 10 shows the main estimates for the parametric model. Table 14 in the Appendix provides evidence on in-sample model fit that is in line with our simulation results. The results suggest a superior model fit of the LLNB as compared to all parametric models. The parametric model is estimated in several specifications: The model PNB (1), with results in the first column in Table 10 is the negative binomial 2 model version of the linear model in Finkelstein *et al.* (2012), i.e. only the treatment variable and dummies for the survey waves, household size and their interactions are included. PNB (2) with results in the second column of Table 10 uses an extended set of regressors as compared to the initial model, i.e. the variables on gender, education, English preferred language, race variables and a dummy indicating a metropolitan statistical area. Regression (3), called PNB (3), further includes the interaction term of the ITT with income. Finally, model PNB (4) is a flexible specification including the squared version of income as well as all two-way interactions of the treatment variables with the non-technical regressors including income and income squared. More details can be found in the Appendix and in Table 10.

As expected, women visit a doctor more often than men do. Moreover, the demand for doctor visits increases with age. The direct effect of income on the number of doctor visits in the parametric model is virtually equal to zero in models (2) to (4). However, the interaction term of income and the lottery outcome is negative and significant, indicating that the ITT might be larger for individuals with low income. According to the flexible Model (4) the interaction of income squared with the treatment variable is significantly different from zero pointing at an ITT being heterogeneous with respect to income.

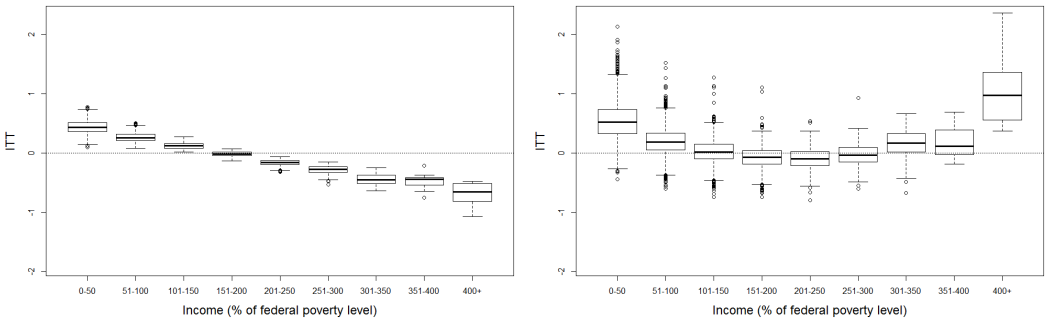
Figure 4 depicts the ITT of being selected in the lottery (i.e. being able to apply for OHP Standard) according to individual income as estimated by the semiparametric model and the parametric regressions (2) to (4). We calculated the ITT for each individual and then plotted the estimated ITTs according to income categories. While the PNB (2) in specification suggests that the ITT is slightly positive and similar in magnitude for virtually all levels of income, the local likelihood estimator reveals a substantially nonlinear effect. Accordingly, the LLNB suggests positive ITTs for the majority of individuals with income below 150% of the federal poverty level. For the majority of observations with intermediate levels of income, the LLNB predicts a negative intent-to-treatment effect (as indicated by a median ITT below zero). The ITT estimates are relatively dispersed for high-income individuals with median effects close to zero up to a level of 350% of the FPL. The PNB (3) incorporates an interaction of the treatment dummy with income leading to a

Figure 4: Boxplots of ITT Effects

(a) LLNB (b) PNB (2)



(c) PNB (3) (d) PNB (4)



The boxplots present the estimated ITTs by the semiparametric and the parametric model in three different specifications. For each individual, the counterfactual difference of the outcome variable is computed that would be expected from a change in treatment status from 0 to 1. ITTs are plotted according to income categories. The boxplot in Figure 4a shows the ITT estimates of the LLNB. Boxplots 4b to 4d present the ITTs as estimated by the parametric models (2) to (4) from Table 10.

(monotonically) decreasing ITT as depicted in Boxplot 4c. The nonlinearity of the ITT in persons' income is confirmed by the results of the flexibly specified parametric Model (4) illustrated in Figure 4d . As the latter includes a squared term of income as well as the two-way interactions of the treatment variable with the (non-technical) covariates, it is sufficiently flexible to reveal the nonlinear ITT.

A plausible explanation for the heterogeneity of the ITT with respect to income might be found in individuals' eligibility. Allen *et al.* (2010) state that the sign-up for the Oregon lottery could not be made conditional on eligibility. The information individuals needed to provide at the registration was kept at a low level in order to reduce potential barriers to enrollment. Before the registered persons were assigned to the treatment or control group, the available information was used in order to exclude ineligible persons (e.g., according to

age or their address). However, this procedure was highly inefficient, due to the limitations of the information (Allen *et al.*, 2010) . As a consequence, truly ineligible persons were assigned to the treatment group. However, only individuals who were truly eligible were later on able to enroll in Medicaid and, thus, experience the expected increase in the number of doctor visits. Allen *et al.* (2010) provide evidence that not meeting the financial criteria (i.e. the income and assets requirements) was one of the major reasons why members of the treatment group did not apply for Medicaid and why, if they submitted an application, their applications were denied. Allen *et al.* (2010) find that only 60% (ca. 18,000) of the treated persons applied for Medicaid and only 30% (ca. 9,000) successfully enrolled in the program. We have no information about the actual treatment status, but observe only the assigned treatment status. As the misclassification of the actual treatment status very likely increases with income, we expect the ITT effect to flatten out with rising income (attenuation bias). Figure 4 appears to illustrate these patterns. The results obtained by the PNB with the flexible specification (4) confirm the nonlinearities of the ITT detected by the semiparametric estimator.

As we have shown in this section, our semiparametric model is not only able to reveal heterogeneous effects in Monte Carlo simulations but also in a real-life application. While the parametric model needs a-priori knowledge about the functional form of the model, the semiparametric model is able to reveal heterogeneous effects without specifying the functional form. Of course, once we have seen the functional form of the heterogeneity in the application using our semiparametric estimator, we were also able to estimate it using a parametric model. However, in real-life applications there may be many potential candidates for heterogeneous effects and hence a completely flexible parametric model may become too involved. In contrast, the LLNB does not require that specification step and detects non-linearities automatically by the data-based bandwidth selection.

5 CONCLUSION

In this study, we proposed a new semiparametric count data model to model health service demand. The major feature of the derived local likelihood estimator is that it addresses the structure of the limited dependent variable (as a count), while maintaining a high degree of flexibility at the same time. It allows abstracting from the linearity assumption embodied in the conditional mean specification of virtually all the usual count data models. The semiparametric estimator enables modeling heterogeneous effects and allows consistent estimation under minimal assumptions. Moreover, our semiparametric estimator explicitly addresses (i) overdispersed and (ii) mixed data, which are

frequently encountered in the estimation of health service demand. The local likelihood negative binomial 2 estimator might be an attractive tool for future research in applied health economics. Our estimator could be employed in the context of robustness checks when there is a concern about the validity of the conditional mean assumption, due to nonlinear or heterogeneous effects. Moreover, heterogeneous patterns might be the object of interest, for instance, in an evaluation of a policy for various subpopulations.

The presented simulation study and empirical application to data from the Oregon Health Insurance Experiment provide encouraging results, based on the predictive power of this semiparametric model as well as on estimation of incremental effects. The results of the simulation study are characterized by a substantial degree of robustness, whereas the local likelihood estimator is found to benefit from larger samples. Furthermore, the performance of the semiparametric estimator has been shown to be superior to parametric and nonparametric alternatives, irrespective of the goodness-of-fit statistic considered. A good performance of the semiparametric estimator can be observed even in presence of excess zeros. As a minor result, the out-of-sample predictive power of the NPCDE has been shown to be inferior to those of the parametric and semiparametric count data models, in particular in the presence of excess zeros. The model-fit results from the Oregon Health Experiment data favor the use of the local likelihood method in estimating the demand for health care. Moreover, the LLNB model was able to reveal a heterogeneity of the intent-to-treat effect with respect to individual income. The detected patterns are in line with economic intuition and the institutional settings, which suggest that the ITT reasonably differs according to individuals' eligibility. Using the parametric NB2 in a linear specification, the heterogeneity in the data would have been missed. If the parametric NB2 model is specified in a sufficiently flexible manner, the parametric results confirm the findings of the semiparametric estimator.

Despite the good results of the local likelihood estimator, there is still room for further improvements. For one thing, the implemented estimator is a local constant estimator, i.e. it does not benefit from the gains regarding bias reduction that can be achieved with higher-order polynomial approximations (Fan *et al.*, 1995). For another, the bandwidths of the local likelihood estimators were obtained by least-squares cross-validation, for convenience of implementation. But, for example, 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 and develop more complex semiparametric count data models: Since the likelihood framework provides explicit expressions for 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. Moreover, the local likelihood framework can be extended to more complex count data settings, for instance, in the presence of hurdles and mixtures.

Acknowledgments

The authors would like to thank the editor, an associate editor, two anonymous referees, and participants at the 25th European Workshop on Econometrics and Health Economics in Odense and seminar participants at the Munich Center for the Economics of Aging for valuable comments.

6 APPENDIX

Note on Implementation

The statistical software used in the simulation study and the empirical application was the 3.2.0 version of R in combination with version 0.99.446 of the R project user interface. The `VGAM` package, version 1.0-0, by Yee (2015) was used for estimating the parametric NB2 model. The NPCDE and the local likelihood estimator were implemented by the 0.60-2 version of the `np` package by Hayfield and Racine (2008).

List of Implemented Models in the Simulation Study

Table 11: List of Implemented Models, Simulation Study

Name	Model	Regressors
LLNB	Local Likelihood Negative Binomial 2 (Semiparametric)	X_{i,c_1}, X_{i,d_1}, X_{i,d_2}
PNB (1)	Negative Binomial 2 (Parametric)	X_{i,c_1}, X_{i,d_1}, X_{i,d_2}
PNB (2)	Negative Binomial 2 (Parametric)	X_{i,c_1}, X_{i,d_1}, X_{i,d_2}, X_{i,c_1}^2, $X_{i,c_1} \cdot X_{i,d_1}$, $X_{i,c_1} \cdot X_{i,d_2}$, $X_{i,d_1} \cdot X_{i,d_2}$
PZNB (1)	Zero-Inflated Negative Binomial 2 (Parametric)	X_{i,c_1}, X_{i,d_1}, X_{i,d_2}
PZNB (2)	Zero-Inflated Negative Binomial 2 (Parametric)	X_{i,c_1}, X_{i,d_1}, X_{i,d_2}, X_{i,c_1}^2, $X_{i,c_1} \cdot X_{i,d_1}$, $X_{i,c_1} \cdot X_{i,d_2}$, $X_{i,d_1} \cdot X_{i,d_2}$
NPCDE	Non-parametric Conditional Density Estimator (Nonparametric)	X_{i,c_1}, X_{i,d_1}, X_{i,d_2}

Variables in the Empirical Application

The variables included in the regression have been subject to transformations to achieve gains in terms of computation time. We build upon the work by Finkelstein *et al.* (2012) who include dummies on the survey waves of the 12 month follow up surveys, dummies on the household size and their interactions since the probability of being assigned to the treatment group varies according to these characteristics (Finkelstein *et al.*, 2012, 1071, 1072). For instance, individuals who registered for the lottery could also register up to two additional persons living in the same household. In the Oregon Health Experiment, treatment assignment was provided to all household members living in the same household as the person selected. Thus, individuals living in larger households were more likely to win access to health insurance. In the original data, the variables *ddddraw_sur_2*, *ddddraw_sur_3*, *ddddraw_sur_4*, *ddddraw_sur_5*, *ddddraw_sur_6*, *ddddraw_sur_7* indicate the survey waves of the 12-months follow up survey. We merged the data from all survey waves following Finkelstein *et al.* (2012). The variables *dddnumhh_li_2* and *dddnumhh_li_3* indicate the number of persons who were additionally registered at the time of lottery sign-up. The exact description of the questionnaire item are listed in the reference "Codebook: Oregon Health Insurance Experiment, Descriptive Variables", available at `http://www.nber.org/oregon/index.html`. The variables *ddddraXnum_2_2*, *ddddraXnum_2_3*, *ddddraXnum_3_2*, *ddddraXnum_3_3*, *ddddraXnum_4_2*, *ddddraXnum_5_2*, *ddddraXnum_6_2*, and *ddddraXnum_7_2* are interactions of the survey and household size indicators and constructed as $ddddraXnum\_j\_i = ddddraw\_sur\_j \cdot dddnumhh\_li\_i$ for $j \in \{1, 2, ..., 7\}$ and $i \in \{2, 3\}$. In order to save computation time we merged the dummy variables on the survey waves *ddddraw_sur_2*, *ddddraw_sur_3*, *ddddraw_sur_4*, *ddddraw_sur_5*, *ddddraw_sur_6*, *ddddraw_sur_7* to a combined categorical variable *ddddraw_sur_comb*. Analogously, *dddraXnum_2_2*, *dddraXnum_2_3*, *dddraXnum_3_2*, *dddraXnum_3_3*, *dddraXnum_4_2*, *dddraXnum_5_2*, *dddraXnum_6_2*, and *dddraXnum_7_2* are combined to one categorical variable *ddddraXnum_comb*. The transformations imposed leave us with a set of 16 instead of 28 variables and hence allow us to substantially reduce the computation time for bandwidth selection.

List of Implemented Models in the Application Study

An overview on the implemented models is presented in Table 12.

Table 12: List of Implemented Models, Empirical Application

Name	Model	Description of Variables included
LLNB	Local Likelihood NB 2 (Semiparametric)	Treatment, survey wave, household size, interactions of survey wave and household size, female, income, age, education dummies (high school/GED, voc. training/2yr-degree, 4-yr degree or more), English preferred language, metropolitan statistical area, race (white, black, hispanic)
PNB (1)	NB 2 (Parametric)	Treatment, survey wave, household size, interactions of survey wave and household size, cf. Finkelstein *et al.* (2012)
PNB (2)	NB 2 (Parametric)	Treatment, survey wave, household size, interactions of survey wave and household size, female, income, age, education dummies (high school/GED, voc. training/2yr-degree, 4-yr degree or more), English preferred language, metropolitan statistical area, race (white, black, hispanic)
PNB (3)	NB 2 (Parametric)	As in PNB (2), plus interaction of treatment with income
PNB (4)	NB 2 (Parametric)	As in PNB (2), plus income squared, and all two-way interactions of the treatment dummy with the covariates (including income and income squared)

Bandwidth Selection in the Empirical Application

Table 13 lists the bandwidths as computed for the LLNB by least-squares cross validation.

Table 13: Bandwidths for Kernel Estimators

Number of Doctor Visits		LLNB
Treatment	ordered	0.0862
ddddraw_sur_comb	factor	0.6625
dddnumhh_li_2	factor	0.1640
dddnumhh_li_3	factor	0.0000
ddddraXnum_comb	factor	0.8889
Female	factor	0.0029
Income (% of FPL)	continuous	44.1640
Age	continuous	7.8395
Educ. HS diploma or GED	factor	0.1235
Educ. Voc. Training or 2-yr degr.	factor	0.4624
Educ. 4-yr College or more	factor	0.5000
English preferred language	factor	0.0511
ZIP	factor	0.5000
White	factor	0.2704
Black	factor	0.5000
Hispanic	factor	0.5000

Application to Health Service Demand: Model Fit

Table 14 presents the results on in-sample model fit as obtained from the semiparametric, and the parametric negative binomial 2 model as specified in Regression (2) to (4) in Table 10.

Table 14: Model Fit, in-sample

	MSE	MAE	RMSE
LLNB	6.5730	1.6591	2.5638
PNB (2)	6.8876	1.7043	2.6244
PNB (3)	6.8734	1.7010	2.6217
PNB (4)	6.8508	1.6982	2.6174

7 REFERENCES

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

Allen H, Baicker K, Finkelstein A, Taubman S, Wright BJ, Group OHS, *et al.*. 2010. What the Oregon health study can tell us about expanding Medicaid. *Health Affairs* **29**(8): 1498–1506.

Baicker K, Taubman SL, Allen HL, Bernstein M, Gruber JH, Newhouse JP, Schneider EC, Wright BJ, Zaslavsky AM, Finkelstein AN. 2013. The Oregon Experiment - Effects of Medicaid on Clinical Outcomes. *New England Journal of Medicine* **368**(18): 1713–1722.

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.

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.

Finkelstein A, Taubman S, Wright B, Bernstein M, Gruber J, Newhouse JP, Allen H, Baicker K, the Oregon Health Study Group. 2012. The Oregon Health Insurance Experiment: Evidence from the First Year. *Quarterly Journal of Economics* **127**(3): 1057–1106.

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.

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.

Taubman SL, Allen HL, Wright BJ, Baicker K, Finkelstein AN. 2014. Medicaid Increases Emergency-Departement Use: Evidence from Oregon's Health Insurance Experiment. *Science* **343**(6168): 263–268.

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

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

hche | Hamburg Center for Health Economics

hche Research Paper Series, ISSN 2191-6233 (Print), ISSN 2192-2519 (Internet)

2011/1 Mathias Kifmann and Kerstin Roeder, Premium Subsidies and Social Insurance: Substitutes or Complements? March 2011

2011/2 Oliver Tiemann and Jonas Schreyögg, Changes in Hospital Efficiency after Privatization, June 2011

2011/3 Kathrin Roll, Tom Stargardt and Jonas Schreyögg, Effect of Type of Insurance and Income on Waiting Time for Outpatient Care, July 2011

2012/4 Tom Stargardt, Jonas Schreyögg and Ivan Kondofersky, Measuring the Relationship between Costs and Outcomes: the Example of Acute Myocardial Infarction in German Hospitals, August 2012

2012/5 Vera Hinz, Florian Drevs, Jürgen Wehner, Electronic Word of Mouth about Medical Services, September 2012

2013/6 Mathias Kifmann, Martin Nell, Fairer Systemwettbewerb zwischen gesetzlicher und privater Krankenversicherung, July 2013

2013/7 Mareike Heimeshoff, Jonas Schreyögg, Estimation of a physician practise cost function, August 2013

2014/8 Mathias Kifmann, Luigi Siciliani, Average-cost Pricing and Dynamic Selection Incentives in the Hospital Sector, October 2014

2015/9 Ricarda Milstein, Jonas Schreyögg, A review of pay-for-performance programs in the inpatient sector in OECD countries, December 2015

2016/10 Florian Bleibler, Hans-Helmut König, Cost-effectiveness of intravenous 5 mg zoledronic acid to prevent subsequent clinical fractures in postmenopausal women after hip fracture: a model-based analysis, January 2016

2016/11 Yauheniya Varabyova, Rudolf Blankart, Jonas Schreyögg, Using Nonparametric Conditional Approach to Integrate Quality into Efficiency Analysis: Empirical Evidence from Cardiology Departments, May 2016

2016/12 Christine Blome Ph.D., Prof. Dr. Matthias Augustin, Measuring change in subjective well-being: Methods to quantify recall bias and recalibration response shift, 2016

2016/13 Michael Bahrs, Mathias Schumann, Unlucky to be Young? The Long-Term Effects of School Starting Age on Smoking Behaviour and Health, August 2016

2017/14 Konrad Himmel, Udo Schneider, Ambulatory Care at the End of a Billing Period, March 2017

2017/15 Philipp Bach, Helmut Farbmacher, Martin Spindler, Semiparametric Count Data Modeling with an Application to Health Service Demand, September 2017

The Hamburg Center for Health Economics is a joint center of Universität Hamburg and the University Medical Center Hamburg-Eppendorf (UKE).

hche | Hamburg Center for Health Economics

Esplanade 36
20354 Hamburg
Germany
Tel: +49 (0) 42838-9515/9516
Fax: +49 (0) 42838-8043
Email: info@hche.de
http://www.hche.de
ISSN 2191-6233 (Print)
ISSN 2192-2519 (Internet)

HCHE Research Papers are indexed in RePEc and SSRN.
Papers can be downloaded free of charge from http://www.hche.de.