THE INFLUENCE OF BMI ON MEDICAL COSTS: A PANEL DATA APPROACH*

TONI MORA[a], JOAN GIL[b] & ANTONI SICRAS-MAINAR[c]

[a] *International University of Catalonia and IEB (UB), Barcelona, Spain*
[b] *CAEPS and University of Barcelona (UB), Barcelona, Spain*
[c] *Badalona Serveis Assistencials (BSA), Badalona, Barcelona, Spain*

VERY PRELIMINARY VERSION
DO NOT QUOTE WITHOUT PERMISSION

Abstract

This paper estimates the impact of BMI and obesity on direct medical costs by using a panel of medical and clinical administrative records for the period 2004-2010 for a subpopulation of Spanish individuals. Such marginal impacts are derived from a two-part model of medical costs (i.e., heavily right-skewed with many zeros) under two different approaches: i) the adjusted-heteroscedasticity retransformed OLS estimation on log costs and ii) the GLM panel data regression. All the estimations account for different challenges posed by the data and a dynamic version of the model is also examined.

JEL Classification:
Keywords: BMI and Obesity; Healthcare costs; Panel data; Two-part models.

* The authors wish to thank BSA (Badalona Serveis Assistencials) for having provided us the core dataset to carry-out this investigation and CESCA (Centre for Scientific and Academic Services of Catalonia) for allowing to use its ICT infrastructures to perform the calculations. We are also indebted to the Catalan Health Department for having provided the access to the Population Census data. The authors acknowledge the comments and suggestions received from a previous version of the paper presented at the XXXI Jornadas de Economia de la Salud (Mallorca, May 2011) and the IV Fedea Health Economics Workshop (Madrid, December 2011). Toni Mora and Joan Gil gratefully acknowledge the financial support of 2009-SGR-102 and 2009-SGR-359 grant programmes from the Generalitat of Catalonia, respectively.

1. Introduction

Obesity -the accumulation of excessive fat in the body- is considered a multifactorial chronic disease linked to genetic, perinatal, socioeconomic and other factors. The prevalence of obesity has tripled in Europe in the last two decades and is estimated that 150 million adults and 15 million children and adolescents in the region are obese (Berghöfer et al., 2008). For the EU15 annual deaths attributable to overweight were 7.7%, and ranged from 5.8% in France up to 8.7% in England. Similar trend is observed in the US and Spain is not an exception, with an estimated 37.8% (15.6%) of Spanish adults being overweight (obese) (Aranceta-Bartrina et al., 2005). After the United Kingdom, Spain is the EU country with the highest increases in obesity rates over the last decade (WHO, 2002); and appears to be one of the countries where the impact of obesity on avoidable mortality is the highest, being responsible for approximately 5.5% of total mortality and about 18,000 deaths yearly (Banegas *et al.*, 2003).

The epidemic is a major public health concern as it is a key risk factor for a range of chronic conditions (i.e., hypertension, diabetes, cholesterol, heart disease, stroke, gallbladder disease, biliary calculus, narcolepsy, osteoarthritis, asthma, apnea, dyslipidemia, gout and certain cancers) which tend to reduce the quality of life and ultimately cause death (Alberti et al., 2009; López-Suárez et al., 2008). In addition, a significant share of obese patients tend to suffer mental disorders and social rejection leading to a loss of self-esteem, a particularly sensitive concern in children (Gariepy et al., 2010). As a consequence of its high prevalence and association with multiple chronic illnesses, obesity tends to substantially increase health care resource utilization and costs.

The connexion between obesity and medical costs in the literature on health economics is rooted in the tradition of the Grossman's model (1972) whereby obesity impacts both the demand of health and healthcare services through the depreciation of the stock of health. Empirical evidence documents that obese people tend to reduce the demand of health but increase the demand of healthcare resouces, thus impacting healthcare budgets.

The aim of this paper is to estimate the influence of BMI and obesity on direct medical costs (i.e., diagnosis and treatment) taking advantage of a panel sample of medical and clinical administrative records of patients observed in the course of seven consecutive years (2004-2010). This study contributes to the literature on obesity economics by modelling medical costs by means of a two-part model (2PM) using panel data econometrics and

examining two broadly studied econometric strategies for the second part of the 2PM. In addition, the estimations incorporate a set of different econometric challenges posed by using panel data analysis. The paper is organised as follows: Section 2 presents the related literature; Section 3 describes the empirical strategy; Section 4 describes the data; Section 5 presents the results, Section 6 discusses the main policy implications of the findings and Section 7 concludes.

2. Related Literature

There exists a considerable body of literature which quantifies the magnitude of the healthcare expenditures associated with the obesity epidemics. Following Barret et al. (2011) we can distinguish two different lines of research. On the one hand, a set of studies are concerned with the estimation of annual direct costs of obesity at an aggregate level. Most of them follow an "etiologic fraction approach" considering the most frequent obesity-related diseases (Wolf and Colditz, 1998; Colditz, 1999; Sander and Bergemann, 2003; Vazquez-Sanchez and Alemany, 2002; Müller-Riemenschneider et al., 2008), while others rely on representative sample data (Finkelstein et al., 2004; Arterburn et al., 2005). These studies find that the share of the national health care expenditures attributable to obesity ranges from 5.3 to 7% for the US and from 0.7 to 2.6% in other countries, reaching for instance a 7% of total health care expenditures in Spain.[1] On the other hand, focusing on a lifetime perspective and using medical records there exists a set of investigations aimed to study the impact of BMI categories on the use of resources and direct costs. Most of these works are drawn from US data (Quesenberry et al., 1998; Thompson et al., 2001; Raebel et al., 2004; Finkelstein et al., 2005) and very few from other country contexts (Borg et al., 2005; Kakamura et al., 2007; van Baal et al., 2008).

A part from to these two strands of research, it is worthy to note the paper by Cawley and Meyerhoefer (2011) which is the first, as far as we know, to estimate the (causal) impact of obesity on medical costs (using MEPS 2000-2005 data) by applying health econometrics methods widely employed in the literature to predict healthcare expenditures. Another key element of their paper is the use of IV methods to address problems of endogeneity and reporting error biases.

[1] Within this type of literature another set of papers estimate medical costs and obesity based on survey data (Sturm, 2002; Andreyeva et al., 2004; Von Lengerke et al., 2006).

3. Empirical Method

There is a plethora of investigations in the field of health economics exploring the advantages and drawbacks of the proposed empirical methods to analyse the utilization of healthcare services and the implied medical costs.[2] The (cross-section) datasets used for analysing such healthcare outcomes typically contain a large proportion of zero observations (non-users) as well as a long right-hand tail of individuals who make a heavy use of healthcare services and imply high costs (skewness). Under these characteristics OLS estimation is biased and inefficient. A leading alternative model to analyse these outcomes and deal with such data problems is the well-known "hurdle" or "two-part model", which assumes that the censoring mechanism and the outcome may be modelled using two separate processes or parts (Manning et al., 1981; Duan et al., 1983; Duan et al., 1984). For instance, in explaining individual annual hospital expenses, the first part determines the probability of hospitalization while the second part explains associated hospital expenditures conditional on being hospitalised.[3]

The traditional candidates for modelling the first part in this literature are binary regression models (i.e., probit and logit). However much controversy exists regarding the estimation of the dependent variable in the second part. On the one hand, researchers have proposed the log transformation of costs (also the square root) before OLS estimation in order to accommodate or reduce skewness (Manning et al., 1983; Duan et al., 1983; Leung and Yu, 1996).[4] As nobody is interested in log model results *per se* (e.g., log dollars) such estimates must be retransformed to the original scale, but these retransformations can be problematic due to the impact of, for instance, heteroscedasticity (Manning, 1998). On the other hand, generalised linear models (GLMs) have been more recently proposed as an alternative approach when there are unknown forms of heteroscedasticity (Mullahy, 1998; Manning and Mullahy, 2001; Buntin and Zaslavsky 2004; Manning et al. 2005). These models specify a distribution function (e.g., Gamma, Poisson, Gaussian) that reflects the relationship between the variance and the raw-scale mean functions and a link function that relates the conditional mean of medical costs to the covariates. Interestingly, GLM estimates are on the raw medical

[2] See Jones (2010) for a review of the proposed econometric methods and their comparative performance.
[3] These two distinct processes can be understood under the perspective of a principal-agent model where the decision to contact a physician is made by the patient but the frequency of visits or continuation of treatment is decided by the doctor.
[4] Actually estimates based on logged models are often much more precise and robust than direct analysis of the unlogged original dependent variable (Manning, 1998). They may also reduce (but not eliminate) heteroscedasticity.

cost scale, so there is no need of retransformation. Another advantage is that this approach allows for heteroscedasticity through the choice of the distribution function.

3.1 Two-part model strategy

Based on the previous literature, this paper estimates direct medical costs by means of a 2PM taking into account the panel structure of the data. Interestingly, in our dataset medical costs are zero for 16% of the sample and positive medical costs are very right-skewed. Thus the first part models the probability of incurring in a positive cost ($y_i > 0$) through a random effects (RE) logit or probit binary model of the type,

$$E(y \mid x_{it}\beta, \alpha_i) = \Pr(y_i > 0 \mid x_{it}\beta, \alpha_i) = F(\alpha_i + x_{it}\beta) \qquad (1)$$

where the non-linear function $F(\cdot)$ is the logistic or the standard normal cumulative distribution function, x_{it} are the regressors and α_i is the unobserved time-invariant and individual-specific effect which is normally distributed, $\alpha_i \sim N(0, \sigma_\alpha^2)$. While the second part of the 2PM uses panel data linear methods to predict mean direct medical costs conditional on positive costs. Notice that these two parts are assumed independent and are estimated separately. In particular two specifications are analysed here:

i) First, a RE generalised least squares (GLS) regression of log medical costs (log y) on a set of controls,

$$E(\log y \,|\, y_i > 0, \alpha_i, x_{it}) = x_{it}'\delta + (\alpha_i + \varepsilon_{it}) \qquad (2)$$

where x_{it} are regressors, $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ is the idiosyncratic error term and both errors are independent of x_{it}. Given that the combined error is $u_{it} = \alpha_i + \varepsilon_{it}$ [with $Var(u_{it}) = \sigma_\alpha^2 + \sigma_\varepsilon^2 = \sigma_u^2$ and $Cov(u_{it}, u_{is}) = \sigma_\alpha^2$, $s \neq t$] it follows that the RE model permits serial correlation over time: $\rho_u = Corr(u_{it}, u_{is}) = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2)$ for all $s \neq t$. In this model is assumed that the individual specific effect is uncorrelated with the explanatory variables.

If (combined) residuals from log medical costs in equation (3) are log-normal and homoscedastic, then the retransformation to raw scale medical costs using the exponentiation function is not a serious problem. Problems become more evident when we deviate from these

circumstances. If the error terms of the logged or transformed model are not normally distributed, but are homoscedastic, the usual alternative for the retransformation has been to relay on the Duan's (1983) smearing or retransformation factor applied in several Rand Health Insurance Experiment papers (e.g., Duan et al., 1983, 1984; Manning et al. 1987). In this case the expected value of medical costs in levels conditional on positive costs is,

$$E\left(y \mid y_i > 0, \alpha_i, x_{it}\right) = e^{\left(\hat{\alpha}+\hat{\gamma}x_{it}^{'}\hat{\delta}\right)}\hat{D} \qquad (3)$$

where $\hat{\alpha}$ and $\hat{\delta}$ are consistent parameter estimates of equation (2) and $\hat{D}$ is the smearing factor, that is, the average exponentiated OLS residuals of the logged dependent variable ($\hat{D} = N^{-1}\sum_{i=1}^{N} e^{(\hat{u})}$) where $\hat{u} = \log y - \hat{\alpha} - x_{it}^{'}\hat{\delta}$.[5] As the typical value for the smearing factor lays between 1.5 and 4.0 in healthcare costs applications, the fact of ignoring the retransformation can produce substantial underestimation of mean medical costs.

However, according to Manning (1998) and Mullahy (1998) this strategy is problematic when transformed errors have a heteroscedastic distribution with a variance that depends on regressors in a nontrivial manner (i.e., $Var(u \mid x) = \sigma_u^2 h(x)$, where *h(x)* is some function of the covariates *x* that determines the heteroscedasticity). These authors point out that OLS estimates of $E\left(y \mid y_i > 0, \alpha_i x_{it}\right)$ that ignore the possible dependence of the retransformation factor on regressors and, therefore, use instead the (homoscedastic) smearing factor are likely to yield biased estimates of key parameters of interest like marginal effects or elasticities.

Given the presence of heteroscedasticity -detected by means of the Breusch-Pagan or White test- if it is produced by several covariates and some of them are continuous (i.e., complex heteroscedasticity) one alternative is to assume a parametric structure for the heteroscedastic error term. Here we follow Mullahy (1998) and assume the exponential conditional mean (ECM) specification accounting for the panel structure of the data: $\sigma_u^2 h(x) = e^{(\alpha + x\gamma)}$ which ensures the positivity of the variance function. Therefore, the heteroscedasticity adjusted retransformation of the expected response of medical costs on explanatory variables is,

[5] When errors are lognormally distributed and homoscedastic, u ~ N(0,σ_u^2), then Equation (3) becomes $E\left(y \mid y_i > 0, \alpha_i x_{it}\right) = e^{\left(\hat{\alpha}+\hat{\gamma}x_{it}^{'}\hat{\delta}+0.5\sigma_u^2\right)}$.

$$E\left(y \mid y_i > 0, \alpha_i, x_{it}\right) = e^{\left(\hat{\alpha} + x_{it}^{'}\hat{\delta} + 0.5e^{(\alpha + x_{it}^{'}\hat{\gamma})}\right)} \tag{4}$$

where $\hat{\gamma}$ are the estimated coefficients of the logarithmic regression $\log(u^2) = \alpha + \chi_0 + \gamma_1 x_1 + \gamma_2 x_2 + ... + \gamma_k x_k + e$ and their significance indicate which are the main variables contributing to the heteroscedasticity. Note that equation (4) rests on the assumption of lognormality of residuals.

As long as the purpose is to recover the estimation of the conditional expected direct medical costs in levels for the entire sample under a 2PM setting, we can write,

$$E\left(y \mid \alpha_i, x_{it}\right) = F(\alpha_i + x_{it}^{'}\hat{\beta}) e^{\left(\hat{\alpha} + x_{it}^{'}\hat{\delta} + 0.5e^{(\alpha + x_{it}^{'}\hat{\gamma})}\right)} \tag{5}$$

where F(·) is the logistic or standard normal distribution. Notice that equation (5) adopts the heteroscedasticity adjusted retransformation of the second part of the 2PM.

ii) Second, a GLM panel regression of (positive) direct medical costs on a set of controls,

$$E(y \left| y_i > 0, \alpha_i, x_{it}\right) = \mu_i = f(\alpha_i + x_{it}^{'}\delta) \tag{6}$$

where the link function f(·), the first component of GLMs, relates the conditional mean of costs directly to the covariates. The second component is a distribution function that specifies the relationship between the variance and the conditional mean. Often this is specified as a power function: $Var(y \mid y > 0, \alpha, x) = E(y \mid y > 0, \alpha, x)^{\upsilon} = u^{\upsilon}$. In order to know which specific link (e.g., logarithm, square root, linear) and distribution functions (e.g., Gamma, Poisson or Gaussian) best fit the data we calculated the Pregibon link test and the Park (1966) test, respectively. However, the most popular GLMs specifications in healthcare cost studies include the log link function and the Gamma distribution (e.g., Manning and Mullahy, 2001; Manning et al., 2005). In this case, the expected value of medical costs for the entire sample is computed as,

$$E(y \mid \alpha_i, x_{it}) = F(\alpha_i + x_{it}^{'}\hat{\beta})\, f(\hat{\alpha} + x^{'}\hat{\delta}) \qquad (7)$$

where F(·) is again the logistic or standard normal cumulative distribution functions.

In selecting these two competing approaches to analyse the impact of BMI (or obesity categories) on mean medical costs, we are aware of the advantages and drawbacks of these two methods. For instance, GLM modelling is recommended, instead of log estimation with retransformation, when complex heteroscedasticity is present and residuals are not lognormally distributed. However, Manning and Mullahy (2001) point out that GLM estimation suffers substantial precision losses in face of heavy-tailed log scale residuals or the variance function is misspecified (see also Buntin and Zaslavsky, 2004; Baser, 2007). A general finding that seems to emerge from the literature that compare the performance of these two models for positive expenditures (among other methods) in terms of consistency and precision (Manning and Mullaly, 2001; Buntin and Zaslavsky, 2004; Manning et al., 2005; Baser, 2007; Hill and Miller, 2009) is that no method dominates the other and there are important trade-off in terms of precision and bias, mainly when different subgroups of population or type of medical costs are analysed (Hill and Miller, 2009; Jones, 2010). Notwithstanding, the Mihaylova et al. (2010) review of literature confirms that 2PM models have shown better performance.

Finally, given the difficulties to find adequate exclusion restrictions in the data, the usual procedure, when estimating 2PM models, is to assume the same type of regressors in both parts of the equations. Fortunately, we have in our data information on patients' relatives which allows us to construct the binary indicator of living with relatives (value 1) or alone (value 0). This indicator is then used as an exclusion restriction since we assume that living with relatives influences the decision to seek care and, consequently, on incurring in positive healthcare costs (first equation) but it is irrelevant when estimating the amount of medical costs (second equation).

3.2 Marginal and incremental effects in two-part models

The derivation of marginal effects (MEs) and incremental effects (IEs) in non-linear models are much more complicated than in linear regression models (see Hertz, 2010). In this paper we are interested in estimating the ME of the BMI regressor, x_k, and the IE of the obesity

regressor, x_d, on direct medical costs (measured in levels) under a two-part framework using the above specifications for the second part.

When we estimate mean medical costs using the adjusted heteroscedasticity retransformation model, to calculate the ME (IE) of BMI (obesity) we take the partial derivative of equation (5) with respect to x_k (x_d) holding constant the remaining covariates,

$$\frac{\delta E(y \mid \alpha, x)}{\delta x_k} = \left(\frac{\delta F(\alpha + x'\beta)}{\delta x_k} e^{\left(\alpha + x'\delta + 0.5e^{(x'\gamma)}\right)} \right) + \left(\frac{\delta \left(e^{\left(\alpha + x'\delta + 0.5e^{(x'\gamma)}\right)} \right)}{\delta x_k} F(\alpha + x'\beta) \right) \tag{8}$$

Now if we assume that F(·) is the cumulative logistic distribution, $\Lambda(\alpha + x'\beta) = \frac{e^{(\alpha + x'\beta)}}{1 + e^{(\alpha + x'\beta)}}$, then the ME becomes:

$$\begin{aligned} \frac{\delta E(y \mid \alpha, x)}{\delta x_k} = & \left(\beta_k \Lambda\left(\alpha + x'\beta\right)\left[1 - \Lambda\left(\alpha + x'\beta\right)\right] e^{\left(\alpha + x'\delta + 0.5e^{(\alpha + x'\gamma)}\right)} \right) \\ & + \left(\Lambda\left(\alpha + x'\beta\right)\left(e^{\left(\alpha + x'\delta + 0.5e^{(\alpha + x'\gamma)}\right)} \right)\left[\delta_k + 0.5\gamma_k e^{(\alpha + x'\gamma)} \right] \right) \end{aligned} \tag{9a}$$

and the IE of the obesity indicator becomes:

where the first term in equation (9) is the ME of the probability of positive medical costs with respect the BMI and the second term measures the ME of the heteroscedasticity adjusted conditional medical costs on positive values respect the same regressor.

Now if we calculate the ME using the GLM specification of the second part of the two- part model and assume the standard normal cdf for the first part $\Phi(\alpha + x'\beta) = \int_{-\infty}^{\alpha + x'\beta} \phi(z)dz$, then the partial derivative of equation (7) is,

$$\frac{\delta E(y \mid \alpha, x)}{\delta x_k} = \left(\beta_k \phi\left(\alpha + x'\beta\right) f(x'\delta) \right) + \left(\Phi\left(\alpha + x'\beta\right) f'(x'\delta) \right) \tag{10}$$

3.3 Econometric issues

Some econometric challenges posed by our panel data need to be adequately tackled. First, weight and height are not always measured when patients visit their doctors implying that for a subset of individuals their BMI may have a missing value in time *t*. To solve this inconvenience, we restricted the sample to those individuals who had at least one weight and height measurement. Based on this information we were able to infer individuals' BMI for the analysed period 2004-2010. Second, as long as not having weight and height measurement information may induce sample selection we followed Wooldridge's (2005, page 581) proposal to accommodate its impact. That is, we run a robust probit estimation of not having measurement on covariates for each period *t* and then saved the inverse Mills ratios that were later added in the two-part model equations.

Third, another issue is to estimate the models through FE or RE in a panel data context. Although FE should control for unobserved heterogeneity at the individual level, we preferred the RE option. This decision was driven by the unfeasibility to estimate the same FE in the two parts of the two-part model. To our knowledge, no standard procedure can perform this. Therefore, we used RE panel estimation which relies on normality of the errors and that errors are uncorrelated with observed covariates (x_{it}). Fourth, to allow for the possibility that the observed BMI may be correlated with the time-invariant and individual-specific effect (α_i) we parameterised this association.[6] However, we followed here the Mundlak (1978) procedure which uses within-individual means of the BMI rather than separate values for each year. As a consequence, the original set regressors is augmented with the global BMI mean. Fifth, to additionally control for heterogeneity we considered the impact of the previous year BMI in our regressions. Notice that although some endogenous effects might be still present like for instance a health status shock (e.g., accident or a job loss) producing a clear impact on medical spending (on traumatologic or psiquiatric services) we assumed that no other effects at the individual level can be controlled for.

Sixth, as a robustness exercise we specified a dynamic panel regression equation by including medical costs incurred in the previous year as an additional regressor to capture state dependence. To deal with the known problem of the initial-conditions we followed Albouy *et al*. (2010) proposal which modifies the Wooldridge (2005) approach. In fact, these authors proposed to use the generalised residual of a simple model in cross-section at the initial date but taking into account the two-part model framework. The latter can be considered the best available estimation of over or under propensity to consume at the initial

[6] Following Chamberlain (1980), one option could be to assume that $\alpha_i = \alpha' BMI_i + u_i \sim idd\ N(0, \sigma^2)$ where $BMI_i = (BMI_{i1},..,BMI_{iT})$ are the values of the BMI for every year of the panel, and $\alpha = (\alpha_1,....., \alpha_T)$.

date. Seventh, another worrying sample selection issue occurs if during the analysed period some individuals drop-out from the panel owing to death, immigration, incapacity, etc. We found around 4% of total observations induced attrition as a consequence of deceased individuals. The strategy here was to simply include a dummy on the death occurrence rather than including an additional probability of dropping from the panel. Eighth, to control for non-linearity we alternatively modelled the impact of BMI categories (e.g., overweight and obese compared to normal weight) on both equations of the two-part model. Finally, marginal effects were computed manually as a consequence of having transformed data and were conveniently bootstrapped.[7]

4. Data and variables

The data used are observational and longitudinal data based on administrative and medical records of patients followed-up during seven consecutive years in six primary care centres (Apenins-Montigalà, Morera-Pomar, Montgat-Tiana, Nova Lloreda, Progrés-Raval and Marti i Julià) and two reference hospitals (Hospital Municipal de Badalona and Hospital Universitari Germans Trias i Pujol), serving to more than 110,000 inhabitants in the north-eastern area of Barcelona. This population is mostly urban, of lower-middle socioeconomic status and from a predominantly industrial context. Our sample includes patients aged 16+ who had at least one contact with the system from 1st January 2004 to 31st December 2010, and who were assigned to the above-mentioned healthcare centres during the study period. The study also considers those who deceased during the analysed period. However we exclude subjects who were transferred or moved to other centres and patients from other areas or regions.

This dataset incorporates a rich set of information regarding utilization of healthcare resources (number of visits to the GP, specialist and emergency care; hospitalizations and bed days; laboratory, radiology and other diagnostic tests; or consumption of medicines), clinical measurements of weight and height, chronic conditions and other diseases associated with each patient (according to the ICPC-2), functional limitations, date of admission and discharge, type of healthcare professional(s) contacted and motive of the visit. Moreover, the dataset informs on each patient's age, gender, employment status (active/retired), place of birth and habitual residence.

[7] We pleasantly thank Partha Deb for having provided us the Stata codes to perform such calculations.

Owing to a unique identifier, the register data is additionally merged with the Population Census allowing us to incorporate new variables for each patient (e.g., education or marital status) that were not available in the original sample.

4.1 Data on Healthcare Costs

In addition to its longitudinal nature, the dataset used is well suited for our purposes because provides a wide array of information on healthcare costs. Such information incorporates the specific characteristics of the considered primary and hospital healthcare centres and also the degree of development of the available information systems. Apart from this internal source of information, when needed costs were calculated using data from invoices of intermediate products issued by diferent providers and from prices set by the Catalan Health Service.

The computation of healthcare costs follows a two stage procedure: first data on the incurred expenditures (Financial Accounting) are converted into costs (Analytical Accounting), and then costs are allocated and classificed accordingly.[8] Depending on volume of activity, we consider two types of costs: fixed or semi-fixed costs and variable costs. The former include personnel (wages and salaries, indemnifications and social security contributions paid by the healh centre), consumption of goods (intermediate products, health material and instruments), expenditures related to external services (cleaning and laundry), structure (building repair and conservation, clothes, and office material) and management of healthcare centres, according to the Spanish General Accounting Plan for Healthcare Centres. The latter include those costs related to diagnostic and therapeutic tests and pharmaceutical consumption.[9]

As unit of measurement we use the cost per treated patient during the period in which the subject is observed and for all direct cost concepts imputed for the set of diagnosed episodes. Table 1 shows an estimate of the resulting unitary cost rates considered for the years 2004 and 2010. Therefore, total medical costs per patient in each period will be derived as the sum of fixed and semi-fixed costs (i.e., average cost per medical visit multiplied by the number of medical visits) and variable costs (i.e., average cost per test requested multiplied by the number of tests + retail price per package at the time of prescription multiplied by the

[8] Expenditures not directly related to care (e.g. financial spending, losses due to fixed assets, etc.) were excluded from the analysis.

[9] For instance we considered: (i) laboratory tests (haematology, biochemistry, serology and microbiology), (ii) conventional radiology (plain films requests, contrast radiology, ultrasound scans, mammograms and radiographs), (iii) complementary tests (endoscopy, electromyography, spirometry, CT, densitometry, perimetry, stress testing, echocardiography, etc.); iv) pharmaceutical prescriptions (acute, chronic or on demand).

number of prescriptions). Note that in this study we do not account for the computation of 'out-of-pocket payments' paid by the patient or family as they are not registered in the database. Healthcare costs figures were converted to 2010 Euros using the Consumer Price Index (CPI).

[Insert Table 1 around here]

4.2 Other variables

The body mass index (BMI) of each patient, our continuous variable of interest, was calculated as weight (in kilograms) divided by the square of height (in metres) using clinical or measured information. The traditional problems found in reporting this information are not an issue here. Notice that in our sample not all patients are measured when they visit the physician, however there are other individuals with several measurements along the observed period. We also computed the impact of the BMI categories (obesity and overweight) on medical costs by using the WHO classification that distinguishes between normal-weight ($18 \leq BMI \leq 24.9$ kg/m^2), overweight ($25 \leq BMI \leq 29.9$ kg/m^2) and obesity (BMI of ≥ 30 kg/m^2).[10]

We controlled the estimations of medical costs by demographic characteristics of patients such as age and gender but also the status of immigrant since there exists evidence of a different pattern of use and access to healthcare services of the immigrant population. Notice that non-linear age effects were considered after running the modified Hosmer-Lemeshow Test. We also added a set of dummies of whether the individual was the main beneficiary of the public health insurance, Catalan as a regular spoken language and the employment status (active/retired). Regarding the individuals' health conditions affecting medical costs two groups of indicators were employed. On the one hand, we included the Charlson comorbidity index for each treated patient as the individual case-mix index obtained from the 'Adjusted Clinical Groups' (ACG), a patient classification system for iso-consumption of resources.[11] On the other hand we considered the number of medical episodes

[10] Although BMI is the most widely used measure of obesity, it poses several problems. For instance, the BMI does not take into consideration body composition (adiposity vs. lean weight) or body fat distribution. This means it may fail to predict obesity among very muscular individuals and the elderly.

[11] A task force consisting of five professionals (a document administrator, two clinicians and two technical consultants) was set up to convert the ICPC-2 episodes to the International Classification of Diseases (ICD-9-CM). The criteria used varied depending on whether the relationship between the codes is null (one to none), univocal (one to one) or multiple (one to many). The operational algorithm of the Grouper ACG ® Case-Mix

suffered by each patient during the analysed period as a proxy for the individual's health status. The merging with the Population Census allowed us to control medical costs by the distance from the current place of residence to the healthcare centre and the educational level and marital status of the patient.

We have an initial balanced panel dataset containing 687,218 observations for the whole period 2004-2010. However, when we restrict the sample to those who had at least one weight and height measurement, we end up with a final sample containing 436,941 observations, a 63.6% of the original sample size.

5. Results

5.1 Summary statistics

Descriptive statistics for the main set of variables used in the empirical exercise are presented in Tables 2-4. Table 2 shows that the unconditional mean annual medical costs per patient over the period is 769€ (in 2010 Euros) that is much larger than the unconditional median of 315€; i.e., less than half the mean cost in our final sample. The skewness statistic is 5.8 (compared to 0 for symmetric data) and kurtosis is 80.96 (compared to 3 for normal data). Thus, the distribution of costs in levels is very right skewed. As expected, the logarithmic transformation shrinks the range of variation of costs reducing the above mentioned skewness: a mean medical cost of 5.04€ close to the median of 5.76€ and the skewness (kurtosis) statistic falls to -0.97 (2.84).[12]

[Table 2 around here]

In addition, direct medical costs are zero for 16.5% of the sample (71,924 obs.), a non-negligible portion of zeros, while the number of observations with positive medical costs amounts to 365,017. As Table 3 reports the mean positive annual costs per patient rises to 920.47€, being significantly higher for women (967.72€) than for men (861.84€). As

System consists of a series of consecutive steps to obtain the 106 mutually exclusive ACG groups, one for each patient. The application of ACG provides the resource utilization bands (RUB) so that each patient, depending on his/her overall morbidity, is grouped into one of five mutually exclusive categories (1: healthy users or very low morbidity; 2: low morbidity; 3: moderate morbidity; 4: high morbidity; and 5: very high morbidity).

[12] Compared to the initial sample, Table 2 shows that medical costs have increased which would indicate that those patients with no measurement of weight and height when they visit their physician enjoyed a better health status and incurred lower costs.

expected, medical costs increase as patients' age, with a higher Charlson comorbidity index and with terminal patients.

[Table 3 around here]

Finally, Table 4 summarises the mean and standard deviation values of the variables of interests and controls. In our sample, the mean measured BMI over the period 2004-2010 is 26.88, corresponding to a prevalence of obesity (overweight) of 24% (37%). As expected, mean measured BMI is slightly higher for men (26.93) than for women (26.83), although obesity is higher for women (26% vs. 22%) and overweight for men (43% vs. 31%). Notice that women represent 54% of the sample and are slightly older than men (49.9 vs. 48.6 years-old). The mean Charlson comorbidity index is significantly higher among men (0.82 vs. 0.65) althought the mean number of episodes is larger among women (18.3 vs. 13.7). As for labour status, around 66% of the sample are active or labour participating individuals and the share of individuals who drop the sample as a consequence of death is relatively higher among men (4% vs. 2%).

5.2 BMI, obesity and medical costs

The results of our RE panel data estimations based on a 2PM are presented in Tables 5-7. The three tables show the (bootstrapped) estimates of the MEs (IEs) of measured BMI (obesity) on medical costs using different econometric specifications. Accompaning such estimates, we also report measures of goodness of fit and of predictive performance for each model (i.e., the auxiliary R^2, RMSE – Root mean squared error and the MAPE – Mean absolute prediction error). Notice that all these estimations account for a wide list of controls (see Section 4.2), health district dummies and time dummy variables. In addition, as it was previously mentioned each fitted model incorporates the inverse Mills ratio of not having weight and height measurement, the global BMI mean (i.e., the Mundlak's correction procedure), one-year lagged measured BMI, a dummy for the death ocurrence and a dichotomous exclusion restriction. The number of the bootstrap replications has been set to 200.

The first set of results (Table 5) presents the estimation of the ME of (measured) BMI on direct costs in levels (following equation 9) under three different specifications. It is worth to note that the first part predicts the probability of any medical cost assuming a panel data logit model and the second part, for positive costs, specifies an adjusted-heteroscedasticity

retransformed panel OLS estimation on log costs. The Shapiro-Wilk test of normality of (log) residuals rejects the null hyphotesis that the residuals are normally distributed (W=18.13, p-value=0.000). We find evidence of heteroscedasticity when regressing the squared residuals of log costs on the set of covariates (Chi-squared=$1.18*10^6$, p-value=0.000). A variant of the Park test suggests that several covariates contribute to this heteroscedasticity, which justifies the adjustment of the retransformed log costs. According to the first specification of Table 5 we find that one additional unit of BMI causes a (modest) increase of 5.1€ in annual medical costs per patient. A dynamic version of the model is also investigated in which (log) medical costs incurred in the previous year and/or a one-year lagged cost indicator is(are) included into the model (specifications 2 and 3, respectively). Interestingly, while we find roughly similar findings (i.e., a ME of 4.7€ and 5.0€) the auxiliary R^2 (from a regression of actual log costs on the predicted values) notably increases reaching 34.2% and 38.9%, respectively, suggesting a reasonable goodness of fit. Notice that the specification which performs relatively best based on the lowest RSME and MAPE criterion, which measures precision of the predictions, is the second model that includes a lagged measure of log costs.

However, as mentioned above a significant drawback of the log OLS approach is that the retransformation of the estimates back to the original escale requires knowledge of the degree and form of heteroscedasticity. As pointed out by the empirical literature (Hill and Miller, 2009; ……..) such regression models tend to perform poorly in terms of bias and predictive accuracy, making the GLM more attractive for the second part of the two-part model. This alternative approach is additionally favoured by the fact that the Kurtosis index of log residuals from a panel OLS regression of direct medical costs has an average value of 2.9 in the data. Although this is slightly lower than the normal distribution (3) we believe that GLMs should be reasonably efficient under this degree of skewness (Manning and Mullahy 2001).[13]

We thus estimate in Table 6 the ME of (measured) BMI on annual direct medical costs according to equation (10). Notice that the first part specifies a panel data probit model to estimate positive medical costs while the second part uses GLM panel data regression. According to the first specification based on Gamma GLM with log link (widely used in the literature on health care costs)[14] we find that one additional unit of BMI causes an increase of 8.3€ in annual medical costs per patient, a significant higher impact than the one estimated in

[13] Cawley and Meyerhoefer (2011) follow the same strategy for estimating their models.

[14] The Pregibon link test gives an estimated value of $-0.591*10^{-5}$ (p-value=0.000) which is practically 0, suggesting the logarithm as the link function. The Park (1966) test gives a coefficient $\upsilon = 1.79$ (p-value=0.000) which is consistent with a gamma-class distribution.

Table 5. Notice that in our data the GLM model performs much better than the OLS log costs estimation as long as the RMSE and MAPE (auxiliary R^2) measures decrease (increase). A dynamic approach is also examined by including both the level of (log) costs and an indicator of whether costs were incurred in the previous year. The second specification shows a lower marginal impact, almost 7€, on annual medical costs caused by BMI, but a relatively better performance is achieved compared to the non-dynamic specification. In the last two regressions of Table 6 we estimate the ME of BMI by running a GLM with a gamma distribution but allowing flexibility regarding the shape of the conditional mean function. This is estimated through the extended estimating equations (EEE) approach of Basu and Rathouz (2005) where a Box-Cox transformation for the link function is applied.[15] Based on the third specification we find a statistically significant marginal impact on annual costs in levels of 9.1€, which is larger compared to the estimation when the log link function is assumed. Similarly, the dynamic version predicts a significant although slightly lower ME of almost 8.4€. Interestingly, notice that under these two modelling strategies we obtain a better performance in terms of both higher accuracy (i.e., lower values of the RMSE and MAPE) and higher goodness of fit (i.e., larger auxiliary R^2).

6. Discussion

7. Conclusions

[15] We obtain an estimated value of the link parameter $\hat{\lambda} = 0.189$ (p-value=0.000). This estimate is then used to run the GLM with a gamma distribution function.

References

Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, Fruchart JC, James WP, Loria CM, Smith SC Jr; International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; International Association for the Study of Obesity. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation. 2009; 120: 1640-1645.

Albouy, V., Davezies, L., Debrand, T. (2010) Health expenditure models: a comparison using panel data. Economic Modelling, 27, 791-803.

Andreyeva T., Sturm R., Ringel JS., 2004. Moderate and severe obesity have large differences in health care costs. Obes Res 12: 1936-1943.

Aranceta Bartrina J., Serra Majem Ll., Foz Sala, Moreno Esteban B., 2005. Grupo Colaborativo SEEDO. Prevalencia de la obesidad en España. Med. Clin. (Barc.) 125: 460-466.

Arterburn D.E., Maciejewski M.L., Tsevat J., 2005. Impact of morbid obesity on medical expenditures in adults. International J Obes 29: 334-339.

Barrett AM, Colosia AD, Boye KS, Oyelowo O. Burden of obesity: 10-year review of the literature on costs in nine countries. ISPOR 13th Annual International Meeting, May 2008, Toronto, Ontario, Canada.

Basu A., Rathouz P., 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostatistics 6(1): 93-109.

Buntin M.B., Zaslavsky A.M., 2004. Too much ado about two-part models and transformation? Comparing methods of modelling Medicare expenditures. Journal of Health Economics 23: 525-542.

Berghöfer A., Pischon T., Reinhold T., Apovian C.M., Sharma A.M., Willich S.N., 2008. Obesity prevalence from a European perspective: a systematic review. BMC Public Health. 2008; 8: 200.

Borg S., Persson U., Odegaard K., Berglund G., Nilsson J.A., Nilsson P.M., 2005. Obesity, survival, and hospital costs-findings from a screening project in Sweden. Value Health 8: 562-71.

Cawley J., Meyerhoefer C., 2011. The medical care costs of obesity: an instrumental variables approach. Journal of Health Economics (2010), doi: 10.1016/j.jhealeco.2011.10.003

Chamberlain G., 1980. Analysis of covariance with qualitative data. Rev. Econ. Stu. 47: 225-238.

Colditz G.A., 1999. Economic costs of obesity and inactivity. Med. Sci. Sports Exerc. 31 (11 Suppl): S663-S667.

Duan N., 1983. Smearing estimate: a nonparametric retransformation method. J. Amer. Statist. Assoc. 78: 605-610.

Duan N., Manning, W.G., Morris C.N., Newhouse, J.P., 1983. A comparison of alternative models for the demand for medical care. J. Bus Econ. Stat. 1, 115-126.

Duan et al., 1984.

Finkelstein E.A., Fiebelkorn I.C., Wang G., 2004. State level estimates of annual medical expenditures attributable to obesity. Obes. Res. 12: 18-24.

Finkelstein E.A., Fiebelkorn I.C., Wang G., 2005. The costs of obesity among full-time employees. Am. J. Health Promot. 20: 45-51.

Gariepy G., Nitka D., Schmitz N., 2010. The association between obesity and anxiety disorders in the population: a systematic review and meta-analysis. Int. J. Obes. (Lond). 34: 407-419.

Grossman M., 1972. On the concept of health capital and the demand for health. Journ Pol. Eco. 80: 223-255.

Hertz T., 2010. Heteroskedasticity-robust elasticities in logarithmic and two-part models. Applied Economics Letters 17: 225-228.

Hill S., Miller G., 2009. Health expenditure estimation and function form: applications of the Generalised Gamma and Extended Estimating Equations models. Health Economics (forth.).

Jones A. M., Rice N., Bago d'Uva M.T. Balia S., 2007. Apllied Health Economics, Routledge Advanced Texts in Economics and Finance 8, Routledge, UK.

Jones A. M., 2010. Models for Health Care. HEDG Working Paper 10/01.

Leung and Yu (1996)

López Suárez A., Elvira González J., Beltrán Robles M., Alwakil M., Saucedo J.M., Bascuñana Quirell A., Barón Ramos M.A., Fernández Palacín F., 2008. Prevalence of obesity, diabetes, hypertension, hypercholesterolemia and metabolic syndrome in over 50-year-olds in Sanlúcar de Barrameda, Spain. Rev. Esp. Cardiol. 61: 1150-1158.

Manning et al. (1981)

Manning et al. (1983)

Manning W.G., Mullahy J., 2001. Estimating log models: to transform or not to transform?. Journ. Health Econ. 20: 461-494.

Manning W.G., Basu A., Mullahy J., 2005. Generalised modeling approaches to risk adjustment of skewed outcomes data. Journ. Health Econ. 24: 465-488.

Mihaylova et al. (2010)

Mullahy, J., 1998. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. Journ. Health Econ. 17: 247-281.

Müller-Riemenschneider F., Reinhold T., Berghöfer A., Willich SN., 2008, Health-economic burden of obesity in Europe. Eur J. Epidemiol. 23: 499-509.

Mundlak Y., 1978, On the pooling of time series and cross-section data. Econometrica. 46: 69-85.

Nakamura K., Okamura T., Kanda H., Hayakawa T., Okayama A., Ueshima H., 2007, Health Promotion Research Committee of the Shiga National Health Insurance Organizations. Medical costs of obese Japanese: a 10-year follow-up study of National Health Insurance in Shiga, Japan. Eur. J. Public Health. 17(5): 424-429.

Quesenberry Jr C.P., Caan B., Jacobson A., 1998, Obesity, health services use and health care costs among members of a health maintenance organization. Arch. Intern. Med. 158: 466-472.

Raebel M.A., Malone D.C., Conner D.A., Xu S., Porter J.A., Lanty F.A., 2004, Health services use and health care costs of obese and nonobese individuals. Arch. Intern. Med. 164: 2135-2140.

Sander B., Bergemann R., 2003, Economic burden of obesity and its complications in Germany. Eur. J. Health Econ. 4: 248-253.

Sturm R., 2002, The effects of obesity smoking, and drinking on medical problems and costs. Health Aff (Millwood) 21: 245-253.

Thompson D., Brown J.B., Nichols G.A., Elmer P.J., Oster, G., 2001, Body mass index and future healthcare costs: a retrospective cohort study. Obes. Res. 9: 210-218.

van Baal P.H.M., Polder J.J., de Wit G.A., Hoogenveen R.T., Feenstra T.L. et al., 2008, Lifetime medical costs of obesity: prevention no cure for inceasing health expenditure. PloS Med 5(2), e29, (DOI http://dx.doi.org/10.1371/journal.pmed.0050029).

Vázquez-Sánchez R., López Alemany J.M., 2002, Los costes de la obesidad alcanzan el 7% del gasto sanitario. Rev. Esp. Econ. Salud, Sept-Oct 1(3).

Von Lengerke T., Reitmeier P., John J., 2006, Direct medical costs of (severe) obesity: a bottom-up assessment of over-vs. normal weight adults in the KORA-study region (Augsburg, Germany). Gesundheitswesen 68: 110-115.

Wooldridge J.M., 2005, Simple solutions to the initial conditions problem in dynamic, non-linear panel data models with unobserved heterogeneity,. J. Appl. Econometrics, 20: 39-54.

Wolf A.M., Colditz G.A., 1998, Curent stimates of the economics costs of obesity in the United States. Obes. Res. 6: 97-106.

Table 1. Estimates of unitary costs per patient in 2004 and 2010

Healthcare resources		Unitary costs (€) Year 2004	Unitary costs (€) Year 2010
Medical visits:			
	Visit to Medical Primary Care	16.09	24.37
	Visit to Emergency Care	79.49*	123.48
	Hospitalization (per day)	217.03*	337.13
	Visit to Specialist Care	71.30*	110.76
Complementary tests:			
	Laboratory tests	18.33	22.64
	Conventional radiology	14.64	18.79
	Diagnostic/therapeutic tests	21.37	37.76
Pharmaceutical prescriptions		PVP	PVP

Note: Figures for years 2004-2010 are estimated from linear interpolation based on observed data in 2003 and 2009. Figures for the year 2010 are derived using the same growth rates. (*) These figures were estimated using the growth rate experienced by primary care visits during the period 2003-2009. PVP is retail price.
Source: BSA analytical accounts.

Table 2. Mean of Annual Direct Medical Costs per Patient 2004-2010 (Euros year 2010)

	Initial Sample		Final Sample	
	Costs (in Euros)	Log Costs	Costs (in Euros)	Log Costs
Unconditional Mean	550.47	4.00	769.08	5.04
Unconditional Median	138.97	4.94	315.00	5.76
Standard Deviation	1,146.69	2.94	1,321.01	2.56
Skewness	6.60	-0.35	5.8	-0.97
Kurtosis	101.44	1.60	80.96	2.84
N (Number of obs.)	687,218	687,218	436,941	436,941

Table 3. Mean of Positive Annual Direct Medical Costs per Patient 2004-2010 (Euros year 2010)

	Final Sample with Positive Costs		
	Both Genders	Males	Females
Full sample	920.47 (1,393.34)	861.84 (1,384.30)	967.72 (1,398.80)
	By subgroups of the population:		
Age 16-24	320.90 (423.74)	307.90 (410.92)	332.63 (1,354.43)
Age 24-40	390.40 (607.38)	380.78 (664.52)	398.32 (555.83)
Age 40-54	624.72 (852.38)	574.61 (855.90)	664.21 (847.53)
Age 54-65	1,049.15 (1,246.88)	974.56 (1,212.95)	1,113.64 (1,271.99)
Age + 65	1,906.11 (2,079.87)	1,851.67 (2,129.86)	1,945.51 (2,042.05)
Female	967.72 (1,398.80)	-	-
Immigrant status	420.36 (717.71)	392.74 (790.12)	443.42 (650.24)
Active (labour status)	499.67 (688.79)	473.00 (683.09)	522.72 (692.84)
Charlson index (>0)	1,096.07 (1,264.48)	947.93 (1,133.44)	1,223.69 (1,354.43)
Deceased individuals	3,131.10 (4,543.03)	3,153.66 (4,754.65)	3,104.64 (4,283.44)
N (Number of obs.)	365,017	162,891	202,126

Table 4. Descriptive statistics of control variables. Period 2004-2010

	Final Sample		
	Both Genders	**Males**	**Females**
BMI	26.88 (5.12)	26.93 (4.45)	26.83 (5.64)
Obesity	0.24 (0.43)	0.22 (0.41)	0.26 (0.44)
Overweight	0.37 (0.48)	0.43 (0.50)	0.31 (0.46)
Age	49.28 (18.70)	48.62 (18.28)	49.86 (19.03)
Female	0.54 (0.50)	-	-
Immigrant status	0.05 (0.22)	0.05 (0.22)	0.05 (0.22)
Active (labour status)	0.66 (0.47)	0.69 (0.46)	0.64 (0.48)
Charlson comorb. index	0.73 (1.39)	0.82 (1.50)	0.65 (1.28)
Average number episodes	16.18 (11.19)	13.67 (9.59)	18.33 (11.99)
Deceased individuals	0.03 (0.17)	0.04 (0.19)	0.02 (0.15)
N (Number of obs.)	436,941	202,073	234,868

Note: Figures are mean values over 2004-2010. Standard deviations are reported in parenthesis.

Table 5. Bootstrapped Marginal Effects of Measured BMI on Annual Direct Medical Costs (in Euros year 2010): OLS log costs panel data estimation

Two-Part Model	ME of BMI	RMSE	MAPE	Auxiliary R^2
OLS on Log(y) + Heteroskedasticity-adjusted Retransformed Model (N=365,017)	5.063 (0.84)***	413,166	5,667	0.061
OLS on Log(y) + Heteroskedasticity-adjusted Retransformed Model + Lagged Costs (N=310,780)	4.732 (0.99)***	335,540	2,340	0.342
OLS on Log(y) + Heteroskedasticity-adjusted Retransformed Model + Lagged Costs + Lagged Costbin (N=310,780)	5.003 (0.95)***	388,302	2,693	0.389

Notes: Auxiliary R^2 denotes the R-squared from a regression of actual costs on the predicted values; RMSE denotes the root mean squared error; MAPE is the mean absolute prediction error. Estimations account for an extensive list of covariates, health district dummies and time dummy variables. In addition, all regressions contain one-year lagged measured BMI, Mundlak's correction and a dichotomous exclusion restriction.

Table 6. Bootstrapped Marginal Effects of Measured BMI on Annual Direct Medical Costs (in Euros year 2010): GLM panel data estimation

Two-Part Model	ME of BMI	RMSE	MAPE	Auxiliary R^2
GLM- Log link + Gamma distr. (N=365,017)	8.251 (1.54)***	307,422	1,986	0.432
GLM- Log link + Gamma dist. + Lagged Costs & Costbin (N=310,780)	6.955 (1.19)***	260,577	2,172	0.582
GLM- link (power 0.189) + Gamma distr. (N=365,017)	9.101 (1.62)***	295,885	1,883	0.541
GLM- link (power 0.189) + Gamma distr. + Lagged Costs & Costbin (N=310,780)	8.355 (1.52)***	256,683	2,019	0.624
GLM- Log link + Poisson dist.	3.972 (1.94)***	294,615	1,738	0.567

Notes: Auxiliary R^2 denotes the R-squared from a regression of actual costs on the predicted values; RMSE denotes the root mean squared error; MAPE is the mean absolute prediction error. Estimations account for an extensive list of covariates, health district dummies and time dummy variables. In addition, all regressions contain one-year lagged measured BMI, Mundlak's correction and a dichotomous exclusion restriction.

Table 7. Bootstrapped Marginal Effects of Measured Obesity on Direct Medical Costs (2004-2010): GLM procedure

Econometric Modelling	ME of Obesity	RMSE	MAPE	Auxiliary R^2
(1) Standard normal cdf (2) GLM- Log link + Gamma dist.	16.169 (5.54)***	307,394	1,985	0.433

Note: All regressions contain lagged-measured BMI, Mundlak's correction and a dichotomous exclusion restriction. Estimations account for an extensive list of covariates, health district dummies and time dummy variables. N in the 2nd part = 310,780 whereas for the imputed value is 405,872.