The Effect of Chronic Diseases on Household Health Care Expenditures: Evidence from the LSMS for Brazil, India and Russia using Generalized Linear Models

Dele O. Abegunde[†], Anderson E. Stanciole[‡*]
[†] *World Health Organization*
[‡] *Department of Economics and Related Studies, University of York*

May 23, 2005

Abstract

Chronic diseases are a growing issue that affects developing countries. This paper explores the effect of chronic diseases on households' expenditures using data from the Living Standards Measurement Studies (LSMS) of the World Bank. The data consists of household micro data for Brazil (1996), India (1997) and Russia (from 1996 to 2002) and we use Generalized Linear Models in the context of two-part models to estimate the marginal effects of chronic diseases. We find that chronic diseases increase both the probability of having positive expenditures and the level of expenditures conditional on having any.

1 Introduction

Chronic diseases have been traditionally associated with developed countries. In recent times, however, the advancement of the epidemiologic transition in developing countries has led this kind of health problems to become an additional challenge that faces such countries. Factors such as dietary changes, increased use of tobacco and declining physical activity, associated with the decline of infant mortality and of the prevalence of communicable diseases, have led to a vigorous increase in the importance of chronic diseases for developing countries. Such change is so vigorous that some authors consider that developing countries are in fact experiencing an epidemic of chronic diseases (Greenberg, Raymond and Leeder 2005).

*Corresponding author. Tel.: +44 1904 622870; email: aes118@york.ac.uk.

Therefore, it is legitimate to consider how the increase in the incidence of chronic diseases may affect the economy and the society in developing countries. In the present paper we evaluate a particular consequence of the spread of chronic diseases, namely the burden that it imposes on household expenditures. Our main objective here is to measure the contribution of chronic diseases to increases in total household expenditures. We employ micro data from the Living Standards Measurement Study (LSMS) that have been performed by the World Bank in several countries over many years. We focus on three major countries (Brazil, India and Russia) which are particularly subject to the prevalence of chronic diseases and for which there was data available.

Two characteristics of health care expenditures make it very difficult to model it. First, they typically present a nontrivial proportion of observations equal to zero. During a given period of time, many households or individuals, according to unit of observation, do not need medical or specialized attention and therefore do not incur in any expenditure at all. Second, among those who actually spend, the probability distribution tends to be very highly skewed, with a great asymmetry with respect to the amount spent, and heteroskedastic. Commonly, most households with strictly positive expenditures spend relatively modest amounts. However, there are very small proportions of households with such a high level of expenditure that actually affect the aggregate average. Consider for instance high cost procedures like organ transplants.

In order to deal with these difficulties, health economists usually resort to the use of two part models (2PM) (Blough, Madden and Hornbrook 1999). The first part includes all the observations and models the probability of incurring in any expenditure. The use of any medical care is coded into a binary variable, which is then estimated using models like Probit or Logit.

The second part includes only the observations with strictly positive expenditures and then applies a transformation on the dependent variable. The usual transformation consists on modelling the logarithm of medical expenditures. This transformation reduces the skewness of the cumulative distribution and stabilizes the variance over the observations, inducing the distribution to approximate the Normal. Hence, in principle, it should allow the use of Ordinary Least Squares without any additional complication.

The major difficult associated with this kind of model shows up when one tries to retransform the dependent variable to its original scale. As several authors have remarked ((Manning 1998); (Mullahy 1998); (Manning and Mullahy 2001)), the retransformation from the log-scale to the original scale is affected by the variance of the log-scale. Therefore, it requires the use of a retransformation parameter, which is called smearing estimator. (Duan 1983) provides a method for obtaining a consistent estimate of the smearing factor using the estimated residuals of the transformed model. However, as noted by (Deb and Burgess 2003, p. 5), such method is "not robust to heteroskedasticity in the transformed scale".

In order to avoid complications associated with the retransformation problem, we have opted to employ generalized linear models (GLM) (McCullagh and Nelder 1983). The advantages of this approach are that it imposes "minimal assumptions and obviates the need to transform the data; rather it represents a reparameterization of the model that retains the original scale (...) of the response" (Blough et al. 1999, p. 154).

The remainder of this paper is organized as follows. The next session depicts the application of GLM in the context of 2PM and describes the strategy that was used to select the most appropriate functional form, with special reference to the specification of the link function. We then describe the data sources and present some preliminary results.

2 Econometric model: 2PM using GLM

Assume that y represents health care expenditures and $\mathbf{x}$ is the vector of covariates that one believes to affect y; 2PM make use of fundamental probability to decompose the determination of health care expenditures into two components:

$$E\left[y \mid \mathbf{x}\right] = P\left[y > 0 \mid \mathbf{x}\right] \cdot E\left[y \mid y > 0, \mathbf{x}\right] \tag{1}$$

Thus, in this specification, one would first estimate the probability of incurring expenses. Next one would compute the expected disbursement for those with positive levels of expenditures. The overall expected expenditure would be given by the product of the two estimates.

The first part, also referred to as the hurdle component, models the probability of incurring in any expenditure during a given period of time. It may be estimated using any standard model of binary choice, like the Probit and Logit models.

By its turn, the second part of the model considers only those observations with strictly positive expenditures. One cannot proceed by simply applying OLS to this subset of observations, since doing so would result in counter-intuitive estimates. In other words, in this case if one merely applies the OLS method (that by construction is defined to the whole of real numbers) the estimated values of expenditure might turn out to be negative, which is logically impossible. Another shortcoming of OLS is that, as it would not account for the possibility of heteroskedasticity in the distribution of expenditures, the estimates would be biased.

One possible and very popular alternative consists on imposing a transformation on the dependent variable, such as taking the logarithm of the expenditures. This transformation smoothes the distribution of expenditures, forcing it to approximate a Normal distribution. In this case, the estimated model would be:

$$E\left[\log(y_i) \mid y_i > 0, \mathbf{x}\right] = x_i \beta \tag{2}$$

In principle, one could then retransform the estimated values fitted by Eq. (2) in order to recover the estimates for y in the original scale. There is an additional complication in that retransformation, however. As shown by (Mullahy

1998, p. 252), according to Jensen's inequality, if one simply takes the exponential of the fitted values on the logarithmic scale:

$$E\left[\exp(\log(y_i)) \mid y_i > 0, \mathbf{x}\right] = E\left[y_i \mid y_i > 0, \mathbf{x}\right] > \exp(E\left[\log(y_i) \mid y_i > 0, \mathbf{x}\right]) \quad (3)$$

This implies a retransformation error that calls for the use of an additional component to compensate for it. (Duan 1983) proposed the use of a smearing estimator that would get round the retransformation error represented in Eq. (3). As suggested by (Deb and Burgess 2003, p. 5), the smearing factor ($\rho(\boldsymbol{x}_i)$) may be computed in the following manner:

$$\rho(\boldsymbol{x}_i) = \frac{1}{N}\sum_{i=1}^{N} \exp\left[\log(y_i) - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right] \quad (4)$$

Hence, $\rho(\boldsymbol{x}_i)$ depends on the deviations on the logarithmic scale. It is then used to recover the estimates for health expenditures in the original scale

$$\hat{y}_i = \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) \cdot \rho(\boldsymbol{x}_i) \quad (5)$$

Also according to (Deb and Burgess 2003, p. 5), however, the smearing estimator is "not robust to heteroskedasticity in the transformed scale".

2.1 Generalized linear models

We opt for using GLMs, since this class of models solves the problem of the skewness on the distribution of expenditures and at the same time also avoids the retransformation problem that results from the use of simple log-linear models.

GLMs are an extension of the classical linear models ((McCullagh and Nelder 1983), (Blough et al. 1999)). Therefore, following (McCullagh and Nelder 1983), in order to simplify the exposition we start with the structure of the classical linear models and subsequently we generalize it to the case of GLM.

Consider that the realization of a random variable Y_i is given by a vector of observations $y_i \sim N(\mu_i, \sigma_i^2)$, where the index i denotes the individual observations. The realization of y_i is governed by two components: the systematic and the random components (Nelder and Wedderburn 1972). The systematic component corresponds to the variation that is explained by covariate variables and is given by μ_i, while the random component is given by the residual variation that cannot be explained by the covariates. Thus μ_i may be defined as a function of the p covariates $x_{i1}, x_{i2}, \dots x_{ip}$:

$$E(y_i) = \mu_i = \sum_{1}^{p} \beta_j x_{ij} \quad (6)$$

The classical linear model is thus characterized by the following configuration, where the index i is suppressed by resorting to the matrix notation:

$$\begin{aligned} \boldsymbol{y} &\sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \\ E(\boldsymbol{y}) &= \boldsymbol{\mu} \\ \boldsymbol{\mu} &= \boldsymbol{x\beta} \end{aligned} \tag{7}$$

2.1.1 The generalization

GLMs extend the classical linear models by relaxing the assumptions behind (7). Thus, they are based on three components ((McCullagh and Nelder 1983), (Blough et al. 1999)):

1. The random component.

The independent variable is assumed to follow a probability distribution from the exponential family (see below for list of possible distributions). This implies that there is a relationship between the variance and the mean and this relationship is given by the variance function:

$$var(y_i) = \boldsymbol{\sigma}_i^2 = \phi V(\mu_i) \tag{8}$$

Thus, the specification of a distributional family implies a given relationship between the variance and the mean. For instance, "the variance function for the Gaussian distribution is 1, because variance and mean are independent, whereas for the skewed Gamma and Inverse Gaussian distributions the variance functions are quadratic ($var(y_i) \propto \mu_i^2$) and cubic ($var(y_i) \propto \mu_i^3$), respectively" (Barber and Thompson 2004, p. 198.).

2. The systematic component.

It is assumed the existence of a linear predictor η_i based on the covariates $x_{i1}, x_{i2}, \ldots x_{ip}$:

$$\eta_i = \sum_{1}^{p} \beta_j x_{ij} \tag{9}$$

3. The link function $g(\mu_i)$.

The random and systematic components are assumed to be related by a linear function $g(\mu_i)$. This function, which is usually called link function, in fact determines how the mean of the independent variable (μ_i) is related to the linear predictor (η_i)

$$g(\mu_i) = \eta_i \tag{10}$$

It is clear that the classical linear model is a special case of GLM where the probability distribution of the response variable (component 1) is Normal or

Gaussian ($y_i \sim N(\mu_i, \boldsymbol{\sigma}_i^2)$) and the link function (component 3) is the identity ((Nelder and Wedderburn 1972), (McCullagh and Nelder 1983)).

Table 1 below (from (Blough et al. 1999)) shows possible exponential distribution families, their natural link functions and the associated variance function:

Table 1
Exponential families of distributions

Distribution	Natural link function	Variance function
Gaussian	μ	1
Bernoulli	$\log\left(\frac{\mu}{1-\mu}\right)$	$\mu(1-\mu)$
Binomial	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\mu(1-\mu)}{n}$
Poisson	$\log(\mu)$	μ
Gamma	$\frac{1}{\mu}$	μ^2
Inverse Gaussian	$\frac{1}{\mu^2}$	μ^3
Quasi	$g(\mu)$	$V(\mu)$

Source: (Blough et al. 1999, p. 158.)

3 Data description

We focus on three countries: Brazil, India and Russia. The data comes from the Life Standards Measurement Study (LSMS) household surveys, which consist on a series of surveys performed by the World Bank in developing countries. The first LSMS was conducted in 1985 in Côte d'Ivoire and since then the surveys have been performed in about 30 different countries. According to the World Bank, "the purpose of these surveys was to gather data that would allow the detailed study of household behavior and several aspects of living" (Grosh and Glewwe 1995).

The table A.1 in the appendix provides some descriptive statistics for the variables used in the analysis. The sample sizes are 4,940 households for Brazil, 2,249 households for India and, in average, 3,927 households for Russia.

The LSMSs sample designs are based on the collection of data at the household level. For some variables, such as hours worked, the questions are also asked in an individual basis, so it is possible to disaggregate the observations up to the individual level. For most variables, however, such separation is not possible. This is specifically true for the variables on expenditures, which are the main interest for us[1]. Consequently, we are bounded to use the households as the unit of analysis, instead of the individuals.

Another important reason for using the household as the unit of analysis is that health related decisions usually involve the interests of the household as a whole (Hjortsberg 2000). (Sauerborn, Adams and Hein 1996) argue that households develop "strategies to cope with the economic cost of illness". (Townsend 1995) and (Gertler and Gruber 2002) have shown that households

[1] See (Deaton 1997) for a comprehensive discussion of the analysis of household surveys.

insure consumption against illness by relying on intrahousehold compensations on the labour supply. For instance, if the breadwinner becomes sick, other members of the household are forced either to start to work or to increase the amount of work currently performed.

The questionnaires are composed by several modules, such as labour, education, migration and health, which cover different aspects of the living standards of the household and its individual members. The data collection is based on interviews performed at the households. We have access to information relative both to the household as a whole (e.g., income and expenditure) and to each of the other members in particular (e.g., age, sex, education, health status and hours worked for each member).

Following the structure of the 2PM, we estimate two different types of regression. In the first part, we model the probability of incurring in health care expenditures. To do so, we use the variable *expend*, which is a dummy that captures the fact that the household reported strictly positive health expenditure. It assumes a value of one if the household actually incurred in health expenditure during the relevant period, and zero otherwise. As may be seen from Table A.1 in the appendix, there is a huge variation with respect to this variable. While for Brazil only 56% of the households reported having incurred in health care expenditures, the corresponding figure for India is 98%. In Russia there is a marked increase from 51% to 68% in the last period.

In the second part of the model, the dependent variable is the level of health care expenditures, conditional on having any. This is the variable *healthexp*. It is measured in nominal values of the local currency and is given for the whole household, not for individual members. The amount of health care expenditure for the household is obtained by adding the partial expenditures in different categories of health care goods and services. Among the items that are included are doctors visits, inpatient care, laboratory exams and medicines.

The value of total expenditure is obtained with a similar procedure, by adding the amounts spent in each particular use. In average health care expenditures correspond to approximately 5% of the total household expenditure. They account for 6.4% of total expenditures for Brazil, 9.4% for India, and for Russia they increase from 2.9% in 1996 to 5.0% in 2002.

Two other important variables are *chronic* and *nonchronic*, and they measure, respectively, the number of adults that reported having chronic diseases and non chronic diseases in the household. The list of diseases researched varies from survey to survey. For Brazil, the chronic diseases are heart problem, high blood pressure, diabetes, respiratory problems, digestive problems, gynecological problems, prostate problems, allergy, cancer, bone/muscle/joint problems, neuro-psychiatric problems and high cholesterol. For India they are respiratory problems, heart problem, blood pressure, cataract and other problems affecting sight and permanent disability. Finally, for Russia the chronic diseases included are diabetes, infarction and stroke.

The number of individuals reporting chronic diseases also varies from country to country. In Brazil, there is in average 0.5 cases per household, which correspond to approximately 21% of the individuals in each household. In Russia, this number varies between 0.17 (7% of the members) and 0.19 (9% of the members) cases per household. For India, the incidence seems to be much smaller, at 0.1 cases per household (around 3% of the members of each household).

On the other hand, the definition of non chronic disease is less precise. For Brazil and Russia, the individuals were asked whether they had had any episode of health problem that interfered with their normal activities during the last 30 days before the interview. In the Indian survey, however, we do not have access to such information. In Brazil approximately 0.55 individuals per household reported suffering health problems in the last 30 days before the interview. In Russia the incidence is higher, floating around 0.9.

We also control for the reported health status. The variable *bad* indicates the number of individuals that self-reported as experiencing bad health. It is around 0.12 in Brazil. In Russia one may note a slight tendency for improvement with *bad* starting at 0.42 in 1996 and going down to 0.35 in 2002.

The other covariates may be classified into three categories: demographic variables, risk factors and access indicators. The demographic variables are the mean individual age (*age*), the number of women (*women*), the number of adults in the household, the marital status of the head of the household (*single*) and caste, which is relevant only for India. It is important to remark that only adults (individuals with 18 years or older) are included in the analysis. This is necessary because, as a general rule, children are not affected by chronic diseases, therefore it is desirable to exclude them from the computations.

The risk factors are those variables that are associated with an increased incidence of chronic diseases in a population. We have included them in the cases where they were available. For Brazil, we included the variables *physical* (the number of individuals that engage in physical activity), *overweight* (the number of overweight individuals) and *obese* (the number of overweight individuals), while for Russia we included *overweight*, *obese* and *smoke* (the number of individuals that smoke). The Indian data, however, did not report any information relative to risk factors, so we were unable to include them.

The variables overweight and obese were computed by taking into account the *weight* and *height* of individuals. With such information it is possible to compute the individual body mass index (BMI), by using the following formula:

$$BMI = \frac{weight}{height^2}$$

Individuals with a BMI between 25 and 30 units are classified as overweight, while those with BMI equal or greater than 30 are considered obese. Again, there is an evident difference between the countries, and we note a higher incidence of both overweight and obesity in Russia as compared to Brazil.

The access indicators reflect how easy it is for a household to have access to health care. They basically describe the geographical location of the household. We have included controls for rural and urban areas (as opposed to metropolitan areas). For Brazil we also controlled for the fact that the household was located in the North-East region, which is markedly less developed than the Center-South of the country.

Table 2
Description of covariates

Variable	Description
income	Household monthly income
totalexp	Total monthly expenditure†
healthexp	Health monthly expenditure
nonhealthexp	Non health monthly expenditure
expend	Dummy for households that reported health expenditures
chronic	Number of adults reporting chronic diseases
nonchronic	Number of adults reporting non chronic diseases
bad	Number of adults reporting bad health
physical	Number of adults that engage in physical activity
overweight	Number of overweight adults (BMI $\geq$ 25)
obese	Number of obese adults (BMI $\geq$ 30)
smoke	Number of adults that smoke
adults	Number of adults (18 years or older)
single	Head of household is single
women	Number of women among adults
age	Mean age among adults
age2	Square of age
education	Mean schooling among adults
education2	Square of education
insured	Number of adults covered by health insurance
pay	Number of adults that pay for receiving health care
caste	Dummy for households of middle or backward castes
rural*	Dummy for households in a rural area
urban*	Dummy for households in an urban area
NE*	Dummy for households in the North-East region

*Comparison basis: metropolitan areas of South-East region

4 Preliminary results

4.1 First part of the model - The probability of health expenditures

In this subsection we comment the results of the logistic regression for the determination of the probability of incurring in any health expenditure during the period under consideration. The complete results are reported in the table A.2 in the appendix. The coefficients in the table A.2 represent odds ratios, i.e. they measure the change in the probability of expenditure associated with a marginal change in the respective covariate.

The most important effect to be considered here refers to the health status of the individuals in the household. The existence of non chronic conditions is highly associated with an increased probability of expenditures in Brazil and in all rounds in Russia (this information was not available for India). This result was expected since this variable indicates the number of individuals in the household that experienced health problems that interfered with their normal activities in the 30 days before the interview. For some rounds in Russia, the number of individuals reporting bad health also increases the probability of incurring in health expenditures.

With respect to the risk factors, only the incidences of obesity (positive effect) and of smoking (negative effect) are statistically significant (in some rounds in Russia). The household income does not seem to affect the probability of expenditure. On the other hand, health insurance increases the probability of expenditure for Brazil and for the two first rounds of Russia. For Brazil, it is striking the effect of the variable *pay*. This variable captures the fact that some people would pay to receive care, while others wouldn't (maybe because they use the public sector). Consequently, it causes a massive increase in the probability of expenditure.

The fact that the head of the household is single is consistently associated with higher probabilities of expenditure. This positive association may indicate the existence of some kind of access problem for households headed by couples. On the other hand, the number of women in households increases the probability of expenditures, corroborating a result usually found in the literature.

It is important to note that in the Indian data the mean age of the adults in the household is the only variable that seems to affect the probability of expenditure. Also, the caste to which the household belongs is not relevant in this process.

4.2 Second part of the model - The magnitude of health expenditures

The table A.3 reports the results for the second part of the model. The method used is the GLM with a log link and a Gamma distribution. As remarked above, this model produces estimates of the multiplicative effect of the covariates, since it models the logarithm of the mean of health expenditures as a function of the covariates. Consequently, in order to grasp the marginal effect of a given variable, one should take the exponential of the respective coefficient reported in the table A.3.

For example, the coefficient reported for education for Brazil is 0.19. This means that one additional year of schooling in the household average implies health care expenditure 21% ($= exp(0.19) - 1$) higher. Keeping in mind this word of caution, the results for the second part of the model are qualitatively very similar to the first part.

4.3 The effect of chronic diseases

The table 3 below summarizes the effect of chronic diseases on health care expenditures. The effect on both parts of the model is apparent.

Table 3 - *Marginal effects of chronic diseases*

		Part 1 (Logistic)		Part 2 (GLM)	
Country		Effect	p-value	Effect	p-value
Brazil		132.0%	(0.0000)	66.5%	(0.0000)
India		156.0%	(0.3310)	75.1%	(0.0000)
Russia	Round 6	9.0%	(0.3310)	17.4%	(0.1290)
	Round 7	13.0%	(0.2370)	19.7%	(0.0760)
	Round 8	61.0%	(0.0000)	37.7%	(0.0010)
	Round 9	45.0%	(0.0000)	28.4%	(0.0010)
	Round 10	43.0%	(0.0010)	56.8%	(0.0000)
	Round 11	84.0%	(0.0000)	50.7%	(0.0010)
	Round 12	31.0%	(0.0190)	31.0%	(0.0000)

5 Conclusion

We have used data from the LSMS surveys from the World Bank to examine the effect of chronic diseases on household health care expenditures in developing countries.

We have employed modern methods that are take into account the high proportion of households that do not make health expenditures during a given period and the elevated skewness of expenditures among the remaining households. The main conclusion is that chronic diseases are significantly associated with both a higher probability of incurring in expenditures and with higher expenditures for those that in fact spend.

This result is relevant given the projected increase in the prevalence of chronic diseases in developing countries during the next years. Given the limited reach of public provision of health care and the incomplete coverage of private health insurance in most countries, households have few alternatives for support in the case of being taken hold of by chronic diseases. Therefore, it is desirable that policy makers consider additional tools to address the problem of the burden of chronic diseases upon households.

The next step in the analysis will involve the examination of the effect of chronic diseases on the labour supply of household members. This is important because of the associated effect over the disposable income of the household.

References

Barber, J. and Thompson, S.: 2004, Multiple regression of cost data: use of generalized linear models, *Journal of Health Services Research and Policy* **9**(4), 197–204.

Blough, D., Madden, C. and Hornbrook, M.: 1999, Modeling risk using generalized linear models, *Journal of Health Economics* **18**, 153–171.

Deaton, A.: 1997, *The analysis of household surveys - A microeconometric approach to development policy*, 1st edn, The Johns Hopkins University Press, Baltimore.

Deb, P. and Burgess, J.: 2003, A quasi-experimental comparison of econometric models for health care expenditures, *Unpublished* .

Duan, N.: 1983, Smearing estimate: a nonparametric retransformation method, *Journal of the American Statistical Association* **78**, 605–610.

Gertler, P. and Gruber, J.: 2002, Insuring consumption against illness, *The American Economic Review* **92**, 51–70.

Greenberg, H., Raymond, S. and Leeder, S.: 2005, Cardiovascular disease and global health: Threat and opportunity, *Health Affairs* **Web Exclusive**.

Grosh, M. and Glewwe, P.: 1995, A guide to living standards measurement study surveys and their data sets, *Living Standards Measurement Study Working Papers* **WP 120**.

Hjortsberg, C.: 2000, Determinants of household health care expenditure - the case of zambia, *Unpublished* .

Manning, W.: 1998, The logged dependent variable, heteroscedasticity and the retransformation problem, *Journal of Health Economics* **17**, 283–295.

Manning, W. and Mullahy, J.: 2001, Estimating log models: to transform or not to transform, *Journal of Health Economics* **20**, 461–494.

McCullagh, P. and Nelder, J.: 1983, *Generalized linear models*, 1st edn, Chapman and Hall, London.

Mullahy, J.: 1998, Much ado about two: Reconsidering retransformation and the two-part model in health econometrics, *Journal of Health Economics* **17**, 247–281.

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

Sauerborn, R., Adams, A. and Hein, M.: 1996, Household strategies to cope with the economic costs of illness, *Social Science and Medicine* **43**, 291–301.

Townsend, R.: 1995, Consumption insurance: an evaluation of risk-bearing systems in low-income economies, *Journal of Economic Perspectives* **9**, 83–102.

Table A.1: Means and standard deviations of selected variables

			Russia						
	Brazil	India	Round6	Round7	Round8	Round9	Round10	Round11	Round12
income	812.51	1,796.42	6,197.58	5,815.45	4,684.19	5,418.28	6,368.95	7,205.54	8,057.52
	(1,868.61)	(5,410.72)	(8,475.76)	(7,834.68)	(18,015.85)	(7,548.14)	(7,342.82)	(10,550.98)	(16,048.00)
totmexp	590.55	2,617.79	1,476,505.00	1,661,664.00	10,682.47	4,964.45	6,731.14	8,013.77	10,144.92
	(889.41)	(2,017.24)	(2,086,128.00)	(2,219,472.00)	(498,037.40)	(7,480.89)	(9,383.13)	(10,509.47)	(17,084.79)
healthexp	37.85	245.98	43,037.80	60,218.86	110.71	246.07	340.22	398.28	503.51
	(155.82)	(551.19)	(165,353.80)	(212,088.70)	(351.15)	(821.75)	(1,185.25)	(1,055.90)	(1,463.70)
nonhealthexp	552.70	2,374.88	1,433,468.00	1,601,445.00	10,571.75	4,718.38	6,390.91	7,615.48	9,641.41
	(856.06)	(1,822.56)	(2,076,280.00)	(2,189,388.00)	(498,038.80)	(7,328.81)	(9,164.84)	(10,354.90)	(16,836.89)
expend	0.56	0.98	0.51	0.52	0.57	0.64	0.68	0.70	0.68
	(0.50)	(0.13)	(0.50)	(0.50)	(0.49)	(0.48)	(0.47)	(0.46)	(0.46)
chronic	0.50	0.10	0.17	0.15	0.17	0.18	0.19	0.19	0.18
	(0.74)	(0.32)	(0.46)	(0.43)	(0.47)	(0.47)	(0.49)	(0.47)	(0.47)
nonchronic	0.55	-	0.88	0.87	0.86	0.90	0.93	0.90	0.89
	(0.84)	-	(0.83)	(0.82)	(0.81)	(0.82)	(0.83)	(0.83)	(0.82)
bad	0.12	-	0.42	0.41	0.41	0.39	0.36	0.35	0.35
	(0.43)	-	(0.62)	(0.62)	(0.62)	(0.62)	(0.59)	(0.60)	(0.59)
physical	0.43	-	-	-	-	-	-	-	-
	(0.75)	-	-	-	-	-	-	-	-
overweight	0.59	-	0.67	0.66	0.67	0.63	0.64	0.65	0.65
	(0.73)	-	(0.73)	(0.72)	(0.73)	(0.72)	(0.71)	(0.73)	(0.73)
obese	0.20	-	0.40	0.43	0.42	0.42	0.41	0.44	0.45
	(0.46)	-	(0.60)	(0.62)	(0.62)	(0.61)	(0.61)	(0.63)	(0.64)
smoke	-	-	0.67	0.68	0.68	0.69	0.73	0.74	0.75
	-	-	(0.73)	(0.74)	(0.74)	(0.75)	(0.77)	(0.80)	(0.81)
adults	2.45	3.45	2.11	2.10	2.12	2.10	2.10	2.10	2.11
	(1.17)	(1.90)	(0.90)	(0.91)	(0.93)	(0.93)	(0.96)	(0.96)	(0.99)
single	0.14	-	0.23	0.24	0.24	0.24	0.25	0.26	0.26
	(0.35)	-	(0.42)	(0.42)	(0.43)	(0.43)	(0.44)	(0.44)	(0.44)
women	1.29	0.47	0.60	0.60	0.60	0.61	0.61	0.61	0.61
	(0.76)	(0.14)	(0.25)	(0.26)	(0.26)	(0.26)	(0.27)	(0.27)	(0.27)
age	40.20	25.69	47.45	47.36	47.43	47.19	47.25	47.27	46.92
	(12.81)	(10.92)	(15.52)	(15.62)	(15.68)	(16.02)	(16.10)	(16.11)	(16.20)
education	6.75	2.40	8.50	8.54	8.65	8.77	8.89	8.95	9.00
	(4.02)	(1.29)	(2.10)	(2.05)	(2.01)	(1.95)	(1.85)	(1.77)	(1.73)
caste	-	0.80	-	-	-	-	-	-	-
	-	(0.40)	-	-	-	-	-	-	-
n	*4,938*	*2,249*	*3,593*	*3,561*	*3,618*	*3,775*	*4,169*	*4,352*	*4,419*

Table A.2: First part of the model - Logistic regressions - Dependent variable: expend*

			Russia						
	Brazil	India	Round 6	Round 7	Round 8	Round 9	Round 10	Round 11	Round 12
chronic	2.32	2.56	1.09	1.13	1.61	1.45	1.43	1.84	1.31
	(0.0000)	(0.3310)	(0.3310)	(0.2370)	(0.0000)	(0.0000)	(0.0010)	(0.0000)	(0.0190)
nonchronic	2.26	-	2.53	2.59	2.86	2.53	2.41	3.23	3.06
	(0.0000)	-	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
bad	0.84	-	1.08	1.14	1.10	1.21	1.25	1.07	1.19
	(0.1490)	-	(0.2960)	(0.0920)	(0.2420)	(0.0280)	(0.0140)	(0.5440)	(0.0730)
income	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(0.7570)	(0.4600)	(0.4140)	(0.0040)	(0.0000)	(0.3240)	(0.0000)	(0.4690)	(0.0000)
physical	1.05	-	-	-	-	-	-	-	-
	(0.4360)	-	-	-	-	-	-	-	-
overweight	0.99	-	1.08	1.10	1.02	1.00	1.14	1.03	0.97
	(0.8830)	-	(0.1620)	(0.1270)	(0.7300)	(0.9570)	(0.0450)	(0.6480)	(0.7090)
obese	1.08	-	1.02	1.15	1.07	1.13	1.20	1.20	1.09
	(0.4450)	-	(0.7770)	(0.0440)	(0.3700)	(0.1200)	(0.0210)	(0.0260)	(0.3030)
smoke	-	-	1.01	0.97	0.93	0.99	0.98	0.99	1.03
	-	-	(0.8480)	(0.6870)	(0.2720)	(0.9250)	(0.7200)	(0.8870)	(0.6530)
adults	0.89	1.46	0.78	0.76	0.95	0.81	0.70	0.76	0.61
	(0.0820)	(0.0930)	(0.0010)	(0.0000)	(0.5170)	(0.0320)	(0.0010)	(0.0220)	(0.0000)
single	0.66		0.62	0.53	0.75	0.55	0.68	0.63	0.55
	(0.0030)		(0.0000)	(0.0000)	(0.0340)	(0.0000)	(0.0060)	(0.0010)	(0.0000)
women	1.13	3.40	1.84	1.73	1.53	1.63	2.26	2.39	1.70
	(0.1730)	(0.4760)	(0.0000)	(0.0020)	(0.0130)	(0.0060)	(0.0000)	(0.0000)	(0.0010)
age	0.98	1.08	0.99	0.99	0.96	0.95	0.94	0.98	0.97
	(0.4020)	(0.3100)	(0.5170)	(0.4570)	(0.0170)	(0.0040)	(0.0010)	(0.3260)	(0.1610)
age2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	(0.2660)	(0.3080)	(0.9480)	(0.9130)	(0.0280)	(0.0010)	(0.0040)	(0.4330)	(0.0670)
education	1.16	0.66	0.93	1.07	1.34	1.18	1.00	0.98	1.13
	(0.0010)	(0.3970)	(0.4140)	(0.4570)	(0.0040)	(0.1140)	(0.9830)	(0.8640)	(0.3390)
education2	0.99	1.06	1.01	1.00	0.98	1.00	1.00	1.00	1.00
	(0.0100)	(0.4790)	(0.1780)	(0.8050)	(0.0100)	(0.7600)	(0.7290)	(0.5940)	(0.8860)
insured	1.06	-	1.17	1.09	1.07	1.09	1.09	1.12	1.27
	(0.2390)	-	(0.0000)	(0.0480)	(0.1450)	(0.2120)	(0.2830)	(0.2360)	(0.0280)
pay	17.98	-	-	-	-	-	-	-	-
	(0.0000)	-	-	-	-	-	-	-	-
rural	0.85	-	-	-	-	-	-	-	-
	(0.1630)	-	-	-	-	-	-	-	-
urban	1.33	-	-	-	-	-	-	-	-
	(0.0030)	-	-	-	-	-	-	-	-
NE	1.18	-	-	-	-	-	-	-	-
	(0.0570)	-	-	-	-	-	-	-	-
caste	-	0.95	-	-	-	-	-	-	-
	-	(0.9190)	-	-	-	-	-	-	-
Pseudo R^2	14.5%	4.9%	9.9%	11.0%	13.7%	11.5%	10.6%	14.7%	14.4%
n	*4,938*	*2,239*	*3,577*	*3,454*	*3,257*	*3,124*	*3,217*	*3,214*	*3,180*

Table A.2: First part of the model - Logistic regressions - Dependent variable: expend*

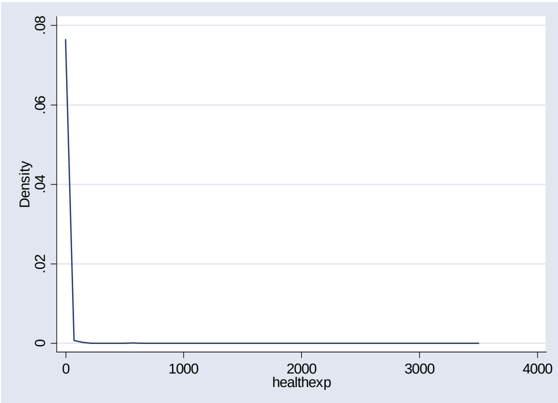
Table A.3: Second part of the model - GLM with log link and gamma distribution - Dependent variable: healthexp*

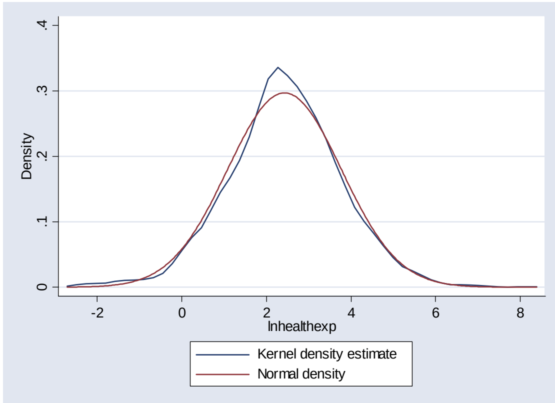
			Russia						
	Brazil	India	Round 6	Round 7	Round 8	Round 9	Round 10	Round 11	Round 12
chronic	0.51	0.56	0.16	0.18	0.32	0.25	0.45	0.41	0.27
	(0.0000)	(0.0000)	(0.1290)	(0.0760)	(0.0010)	(0.0010)	(0.0000)	(0.0010)	(0.0000)
nonchronic	0.38	-	0.79	0.71	0.95	0.68	0.52	0.48	0.68
	(0.0000)	-	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
bad	0.17	-	0.42	0.29	0.27	0.36	0.26	0.27	0.04
	(0.1360)	-	(0.0000)	(0.0040)	(0.0010)	(0.0000)	(0.0060)	(0.0150)	(0.6850)
income	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.5880)	(0.2310)	(0.0010)	(0.0040)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0030)
physical	-0.03	-	-	-	-	-	-	-	-
	(0.6100)	-	-	-	-	-	-	-	-
overweight	0.05	-	0.06	0.09	0.13	0.24	0.11	0.12	0.03
	(0.4210)	-	(0.3940)	(0.2610)	(0.0580)	(0.0020)	(0.1670)	(0.1290)	(0.6560)
obese	0.13	-	0.14	0.11	0.21	0.04	0.17	0.19	0.17
	(0.1460)	-	(0.1130)	(0.1430)	(0.0110)	(0.5880)	(0.0510)	(0.0670)	(0.0620)
smoke	-	-	-0.09	0.05	-0.06	-0.24	-0.02	-0.01	-0.06
	-	-	(0.3210)	(0.5290)	(0.4260)	(0.0040)	(0.7900)	(0.8310)	(0.3570)
adults	-0.10	0.14	-0.25	-0.31	-0.16	0.10	-0.04	-0.04	-0.41
	(0.0480)	(0.0000)	(0.0020)	(0.0000)	(0.0400)	(0.3680)	(0.7450)	(0.7850)	(0.0010)
single	-0.42	-	-0.16	-0.25	-0.44	-0.22	0.24	-0.41	-0.42
	(0.0040)	-	(0.5010)	(0.1920)	(0.0060)	(0.1830)	(0.1970)	(0.0140)	(0.0120)
women	0.09	0.65	0.77	-0.67	0.15	0.08	-0.09	0.39	0.15
	(0.2720)	(0.0570)	(0.0130)	(0.0140)	(0.4720)	(0.7290)	(0.6830)	(0.0540)	(0.4350)
age	0.02	0.00	-0.02	0.01	-0.05	-0.02	0.00	0.00	0.00
	(0.3460)	(0.7750)	(0.4110)	(0.6030)	(0.0210)	(0.4070)	(0.9460)	(0.8400)	(0.9190)
age2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.9520)	(0.7400)	(0.9040)	(0.4050)	(0.0450)	(0.5620)	(0.7000)	(0.9550)	(0.7740)
education	0.19	0.69	0.04	0.03	0.18	-0.13	0.03	0.07	0.20
	(0.0000)	(0.0000)	(0.7550)	(0.7720)	(0.1230)	(0.2630)	(0.7970)	(0.6010)	(0.1060)
education2	-0.01	-0.08	0.00	0.00	-0.01	0.01	0.00	0.00	0.00
	(0.1180)	(0.0000)	(0.5420)	(0.6760)	(0.3520)	(0.0860)	(0.7180)	(0.6130)	(0.8190)
insured	0.17	-	0.18	0.21	0.07	-0.03	-0.05	-0.17	0.17
	(0.0000)	-	(0.0000)	(0.0000)	(0.1870)	(0.6910)	(0.6440)	(0.2490)	(0.1100)
pay	1.52	-	-	-	-	-	-	-	-
	(0.0000)	-	-	-	-	-	-	-	-
rural	-0.02	-	-	-	-	-	-	-	-
	(0.8420)	-	-	-	-	-	-	-	-
urban	0.20	-	-	-	-	-	-	-	-
	(0.0530)	-	-	-	-	-	-	-	-
NE	-0.34	-	-	-	-	-	-	-	-
	(0.0000)	-	-	-	-	-	-	-	-
caste	-	-0.01	-	-	-	-	-	-	-
	-	(0.9560)	-	-	-	-	-	-	-
_cons	1.11	3.45	9.28	9.89	3.48	4.47	4.31	4.39	3.86
	(0.0170)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
AIC	44,938.32	101,990.50	22.8	23.4	10.7	12.4	13.2	13.6	14.1
n	*4,938*	*2,239*	*3,577*	*3,454*	*3,257*	*3,124*	*3,217*	*3,214*	*3,180*

*Numbers between parentheses represent p-values.

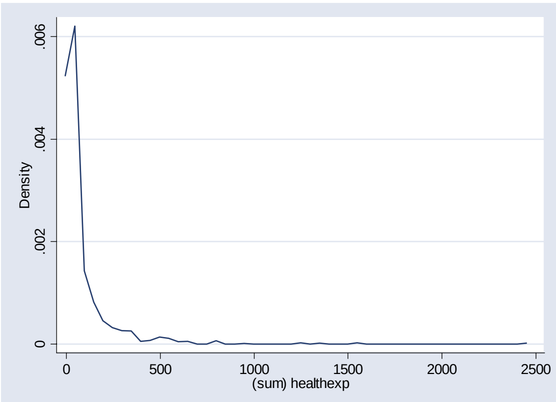
Figures: Cumulative distributions for health expenditures (healthexp) and logarithm of health expenditures (lnhealthexp)

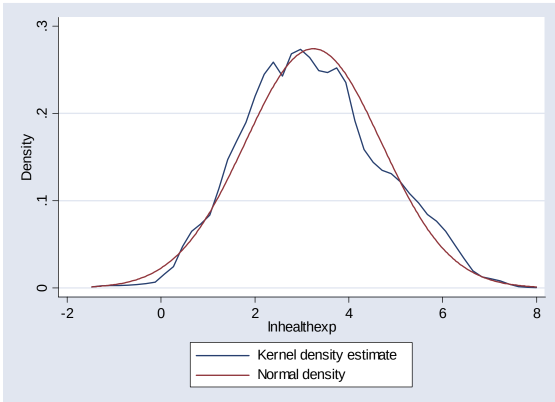
1. Brazil





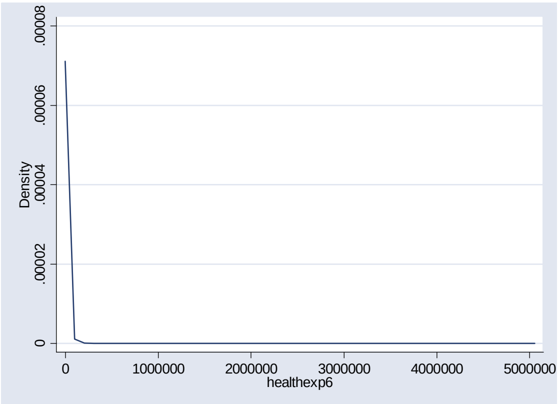
2. India

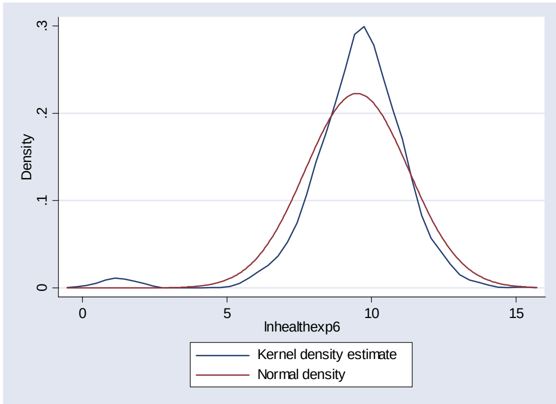




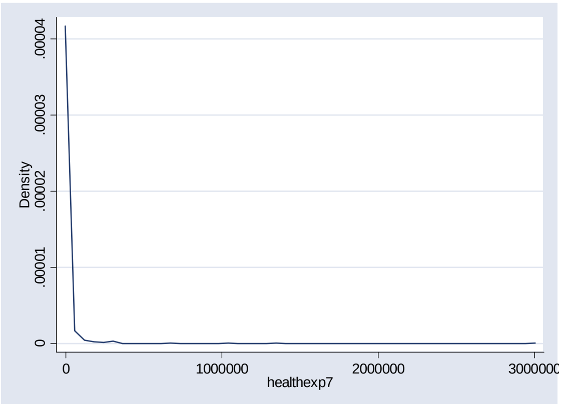
3. Russia

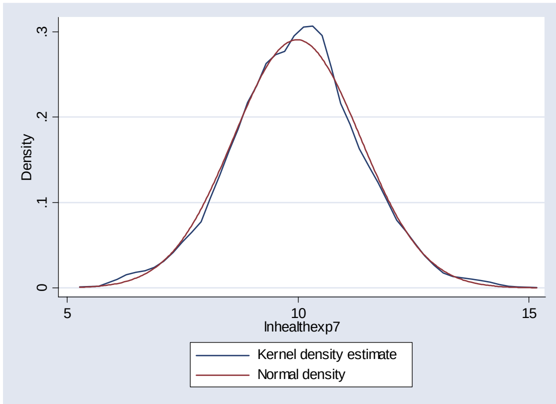
a. Round 6



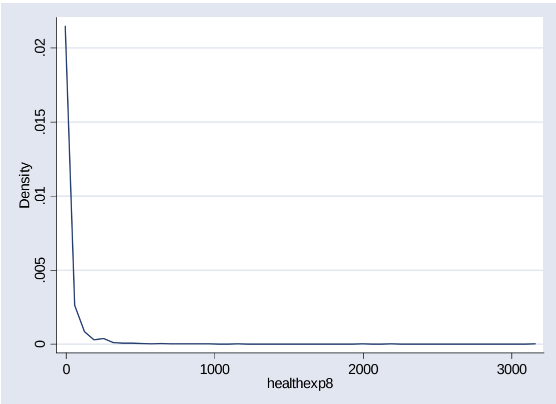


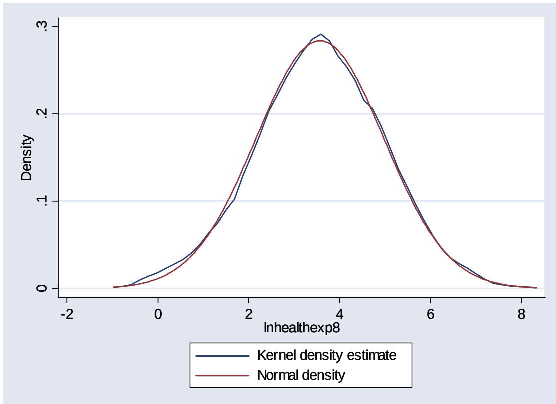
b. Round 7



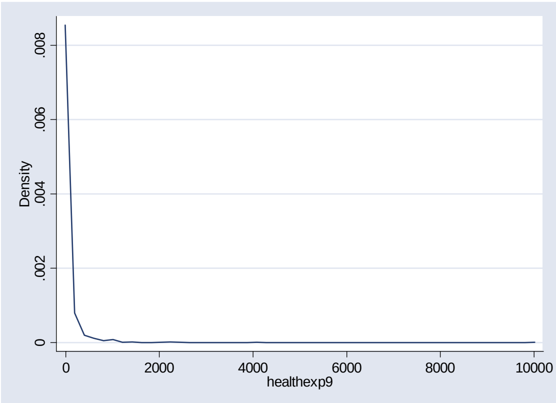


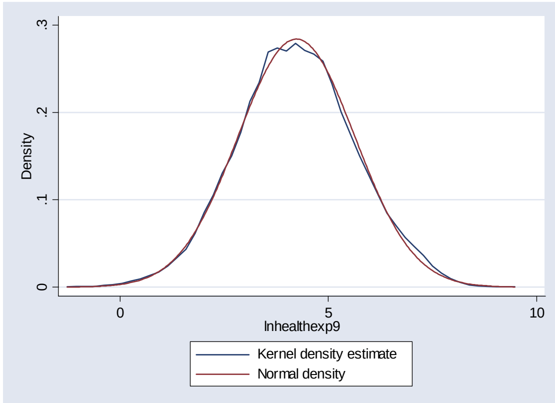
c. Round 8



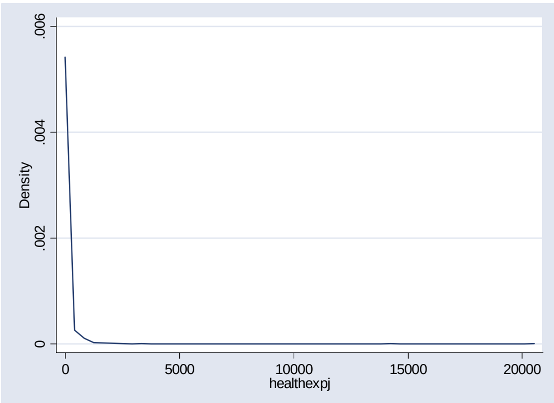


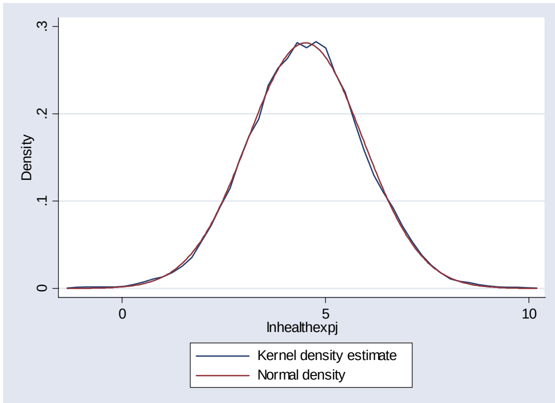
d. Round 9



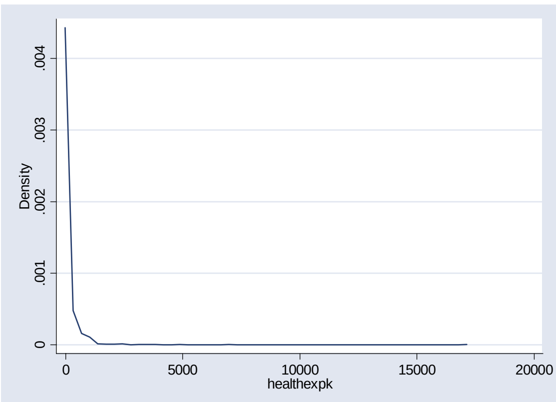


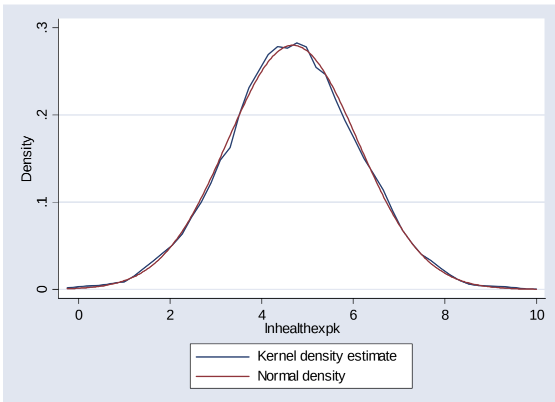
e. Round 10



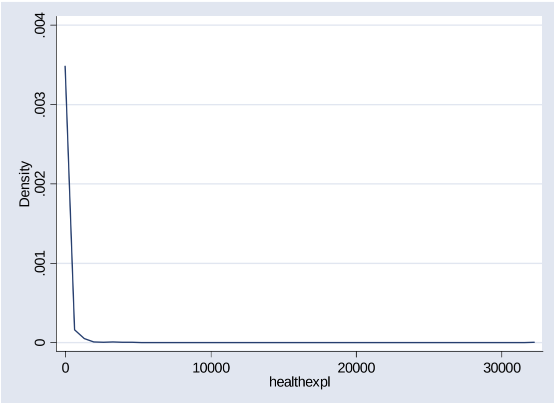


f. Round 11

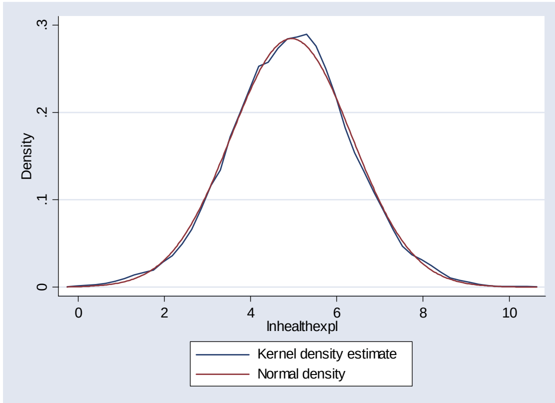




g. Round 12



Density
0
.001
.002
.003
.004
0
10000
20000
30000
healthexpl



Density
0
.1
.2
.3
0
2
4
6
8
10
lnhealthexpl
Kernel density estimate
Normal density