WORKING PAPERS

N° 17-826

July 2017

"Calorie intake and income in China: New evidence using semiparametric modelling with generalized additive models"

Huong Thi Trinh, Christine Thomas-Agnan and Michel Simioni

Toulouse School of Economics

Calorie intake and income in China: New evidence using semiparametric modelling with generalized additive models

TRINH TH.*[1], THOMAS-AGNAN C.[2], and SIMIONI M.[3]

[1]Toulouse School of Economics, INRA, University of Toulouse Capitole, Toulouse, France
[2]Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France
[3]INRA, UMR 1110 MOISA, Montpellier, France

June 30, 2017

Abstract

Recent research on calorie intake and income relationship abounds with parametric models but usually gives inconclusive results. Our paper aims at contributing to this literature by using recent advances in the estimation of generalized additive models with penalized spline regression smoothing (GAM). These semi-parametric models enable mixing parametric and nonparametric functions of explanatory variables and enlarge the distribution of the response variable. The revealed performance test (Racine and Parmeter, 2014), supported by simulation data, shows that GAM models outperform the parametric models. Using data from CHNS in 2006, 2009 and 2011, we find a positive and statistically significant relationship between household calorie intake and household income for the poor. Then the impact of increasing income on calorie consumption slows down for the middle class and the rich. In addition, we find that income-calorie elasticities are generally small, ranging from 0.07 to 0.12.

Keywords: Calorie intake and income, generalized additive models, CHNS data, revealed performance test, cross validation procedure.

*Contact author : trinhthihuong@tmu.edu.vn

1 Introduction

Food consumption patterns of population is a critical problem for both developed and developing countries and its impacts are summarized in Bhargava (2008). While developed countries are dealing with the problem of obesity among children and increasing sedentary lifestyles, developing countries have to handle complex problems of under-nutrition and over-consumption of food. In addition, Popkin (2003) has explored nutrition transition in developing countries and has been concerned with the issue of the burden of nutrition-related non communicable diseases (NR-NCDs), a follow-up stage after receding famine.

There has been an inconclusive debate about whether there exists a strongly significant and positive relationship between household income and calorie demand. Recently, Ogundari and Abdulai (2013) used Meta- regression analysis to examine a total of 40 empirical studies on this issue over the world. The available empirical studies seem to suggest that specifying the relationship between the response variable and the explanatory variables plays an important role. The relationship is potentially nonlinear knowledge, however, among 99 income-calorie elasticities collected in the paper, 86 values are estimated by using parametric models with logarithm or square in order to capture non-linearity. Few papers use semiparametric models to deal with the issue of non linearity (see, for instance Gibson and Rozelle, 2002; Vu, 2009; Nie and Sousa-Poza, 2016; Tian and Yu, 2015). In the case of China, as summarized in Nie and Sousa-Poza (2016), current research appears to validate the view that elasticities vary substantially, even among studies using the same dataset. For example, in the recent study using Chinese Health and Nutrition (CHNS), Nie and Sousa-Poza suggests that "no clear nonlinearity, regardless of whether parametric, nonparametric, or semiparametric approaches are used" while Tian and Yu (2015) claim that "nutrition improvement and dietary change will continue in China but will slow down in the future with further income growth."

In this paper, we discuss semiparametric methods based on penalized spline smoothing as in Ruppert et al. (2004) and generalized additive model as in Wood (2006). The approach is illustrated on the Chinese data from the CHNS survey for the years 2006, 2009 and 2011. A crucial argument is how to specify the relationship between the response variables and the explanatory variables, whether linear or nonlinear. To our knowledge, there are three main arguments that can be advanced to support generalized additive models: a response variable distribution in the exponential family, a non parametric relationship between the expected response and the explanatory

variables and possibility of mixing parametric model and non-parametric functions. We also discuss variable selection issues which can be addressed by stepwise procedures or shrinkage methods discussed in Marra and Wood (2011). The model comparison is based on the cross validation criterion in Racine and Parmeter (2014). In this paper, we also use simulated data from a given data generating process (DGP) and it provides strong evidence that the revealed performance test chooses the true model.
The paper is organized as follows. Section 2 discusses semiparametric regression using penalized spline smoothing. Section 3 describes the idea of cross - validation procedure and simulation results. Section 4 presents the semiparametric model applied to the Chinese data with different choices of distribution for the response variable and compares with the traditional regression approach. The conclusion and several suggestions for policy makers are presented in the final section.

2 Semi-parametric regression with penalized spline smoothing

A generalized additive model (GAM) is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. The theory builds around penalized regression smoothers (see, for example Hastie and Tibshirani, 1990) and Eilers and Marx (1996). In general, the model has the following structure:

$$g(\mathbb{E}(Y_i)) = X_i^{*'}\beta^* + \sum f_k(Z_i) \tag{1}$$

where
Y_i is the response variable following a distribution from the exponential family,
g is a given link function,
$X_i^{*'}$ is row i of the part of the design matrix corresponding to covariates acting linearly on $g(\mathbb{E}(Y))$,
β^* is the parameter corresponding to the linear part of the model,
Z_i is row i of the part of the design matrix corresponding to covariates acting non-linearly on $g(\mathbb{E}(Y))$,
f_k are smooth functions of the covariates acting non-linearly on $\mathbb{E}(Y)$.
The smooth functions can be function of a single covariate as well as interactions between several covariates.
In the generalized additive model, the distribution of the response variable

Y_i belongs to the exponential family with density function given by

$$f_\theta(y) = exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)]$$

where b, a and c are arbitrary functions, ϕ an arbitrary 'scale' parameter and θ is known as the 'canonical parameter' of the distribution. Among the most popular distributions of the exponential family are the Gaussian, Poisson and Gamma.
For fitting the model, we refer the readers to the theoretical papers such that Wood and Augustin (2002), Wood (2006), Wood (2003) and Xiang and Wahba (1996).

3 Cross validation and Simulation results

Model selection and variable selection are very important for the quality of the fit and the predictive power of a model. Several procedures can be used such as cross validation criteria with corresponding theoretical and practical properties and we refer the reader to Zucchini (2000) for a discussion. On theoretical grounds, there is no compelling reason to argue that a semi-parametric model will perform better than a parametric model. To go further, our discussion will point to a cross validation procedure, namely a test for revealed performance, initiated by Racine and Parmeter (2014). The procedure requires splitting the sample into two independent samples of size n_1 (called calibration data) and n_2 (called validation data). The first n_1 observations is fitted on interested models, then we predict the values on the remaining validation data, next we compute average square prediction error (ASPE) since the response values for the evaluation data are given, this presents an estimate of true error. Let $\hat{Y}_V$ be a predictor for Y_V constructed from any specific model estimated using the calibration data, then

$$ASPE = \frac{1}{n_2}\sum_{j=1}^{n_2}(Y_{Vj} - \hat{Y}_{Vj})^2 \tag{2}$$

The ASPE value indicates that one model outperforms the other but the result depends on particular division of the data in two disjoint subsets. To overcome the limitation, it is know that we need to repeat the process many times, say $S = 1000$ times. Thus, the two sample ASPEs generated by the procedure are used to distinguish between the two models. Then, the smaller the ASPE is, the better the predictive power of the model. The revealed

performance test is prominent in the literature on error estimation, since the test approaches the distribution of a model's *true error* (Efron, 1982), and also because the test does not require that compared models be the same type. Along similar lines, we recommend the readers to apply a simple test of differences in means for the two distributions, called the Kruskal-Wallis test and also consider appropriate graphical tools.
Even though there are some simulations on the Racine and Parmeter's work, further evidence supporting the cross validation procedure for our problem, say the income calorie intake relationship, needs to be considered. That is why we conduct three simulation exercises where each model includes a response variable y and an explanatory continuous variables x and a factor variable fac. The simulation models are summarized as follows

- LLdouble is a parametric model of the following form

$$\log(y) = 5 + 0.8\log(x) - 0.007\log^2(x) + 0.4fac + \epsilon,$$

- GauNL is an additive generalized linear model where y follows Gaussian distribution and there is the following link between the response and explanatory variables.

$$\log(\mathbb{E}(y)) = s(x) + 3fac,$$

- NBNL is an additive generalized linear model where y follows Negative Binomial distribution and there is the following link between the response and explanatory variables.

$$\log(\mathbb{E}(y)) = s(x) + 3fac,$$

where, $s(x) = 1 + 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10}$ (the relationship with the response variable is non linear). The factor fac has 2 levels 0 and 1. The number of replications is $n = 3000$. The choice of the LLdouble GDP comes from the fact that it is frequently used in the literature. The function $s(x)$ is complex enough to describe the non-linearity between the covariate and the response variable. In addition, the two data GauNL and NBNL will allow us to measure how the criterion reacts with the different choices of the distribution in the semi-parametric model.
For what follows, we fit each simulated data set with the three models LLdouble, GauNL and NBNL, and then apply the cross validation procedure with $n_1 = 2500$, $n_2 = 500$ and $S = 1000$. The boxplot of the obtained ASPE sample values is reported in Figure 1. Also, Table 1 shows the Kruskal-Wallis

test values to compare the mean of the two ASPE sample values between the pair of fitting models. Visually, for the LLdouble simulated data, the mean of the three samples are quite equal. Along similar lines, for the GauNL and NBNL, the mean of the ASPE sample are well discriminated between parametric model and semi-parametric model. The Kruskal-Wallis test yields a small $p-$value for the pair LLdouble-GauNL and LLdouble-NBNL for the three, indicating a significant difference in the mean. However, between the pair GauNL and NBNL, the $p-$value is large. Thus, we cannot reject the hypothesis that the mean of the two samples are equal. We would like to conclude that in the case of simulated data with linear relationship, there is no difference between applying a parametric model and a semiparametric model. However, in the case a non-linear relationship, the cross validation procedure will choose the true model (here is GauNL data and NBNL data with fitting GauNL and NBNL model).

Figure 1: Boxplot of the ASPE sample

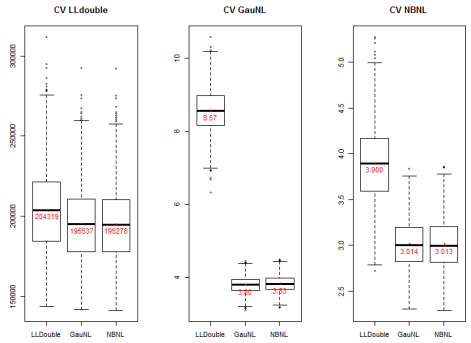


Table 1: $p-$ value of Kruskal-Wallis test

	LLdouble simulation	GauNL simulation	NBNL simulation
LLdouble-GauNL	$4.049e-13$	$< 2.2e-16$	$< 2.2e-16$
LLdouble-NBNL	$8.031e-14$	$< 2.2e-16$	$< 2.2e-16$
GauNL-NBNL	0.793	0.002	0.899

4 The model calorie intake - income in China

4.1 Description of the Chinese data set

The empirical work in this paper uses a data set from the Chinese Health and Nutrition Survey[1]. The survey involves nine provinces displaying variability in geography, economic development, public resources, and health indicators. The number of households and the number of individuals vary each year since many families that migrate from one community to a new one are not followed anymore. Here, we focus on the years 2006, 2009 and 2011 with the purpose of finding empirical results for China. The first reason for this choice is that the relationship between calorie intake and income has changed rapidly with other economical problems in this period. Another reason is that, as far as we know, there are several studies focusing on these same years, (see, for instance Nie and Sousa-Poza, 2016; Tian and Yu, 2015). In these works, the authors apply many different models, either parametric with logarithm or higher order and semiparametric, and for both panel data and cross sectional data. However, the results have been contradictory regarding the relationship between calorie intake and increasing income. While Tian and Yu (2015) find a significant relationship, Nie and Sousa-Poza (2016) conclude that Chinese households are quite successful in maintaining the amount of calorie stable as income vary. After filtering observations, our data set is summarized in Table 3.

These data include calorie intake (expressed in kcal) for three consecutive days for each household and individual, asking all respondents directly about all food consumed inside and outside their home on a 24-hour recall basis. Here, we use total household calorie intake per day (THCC) as a response variable since there are an unequal sharing of calorie intake in the family, with regard to male and female, children and adults.

As covariates, we use household income (HHINC), household characteristics and location characteristics.

The income (HHINC) measures the total income per family which is attributable to nine sources: farming, gardening, livestock/poultry, fishing, business, subsidies, retirement income, non-retirement earnings, and other sources of income, deflated in 2006. There is a rapid increase of income in China. The average household income is 24607 Yuan in 2006, 34829 Yuan in 2009 and 40392 Yuan in 2011.

Our models include five household characteristics variables: household size (HSIZE), availability of safe drinking water (WA), ethnicity of the head

[1]See website `http://www.cpc.unc.edu/projects/china/about/design/datacoll`

of household (ETHNIC), the gender of the head of household (GENDER) as well as the highest attained level of education of the head of family (EDUCH). For location, we consider the urban-rural position of the sites (URBAN) and the province variable (PRO).
The HHINC variable is dealt as a continuous variable and the others are regarded as factors.
For the description about nutrition improvement and dietary change in China with CHNS data, we refer the reader to the paper of Tian and Yu (2015) where the authors propose different indices of nutrition transition.

4.2 Results of model fitting

We apply the traditional parametric model in the literature, say the log-log double model (LLD), and the semiparametric model with generalized additive model, say the GAM, to the Chinese data with several specifications. For implementing GAM in R, we use the package *mgcv* proposed by S.Wood. The choice of distribution for THCC will impact the quality of fit of an model. From the histogram of THCC for each year (see Figure 2), it is clear that the distribution of THCC does not exactly fit with a Gaussian assumption. To go further, we draw the theoretical quantile-quantile plot (or Q-Q plot) for THCC with various distribution in the exponential family including Gamma, Gaussian, Poisson and Negative Binomial. For each year, first we divide the data by decile. Then, with each given distribution, the sample quantile and theoretical quantile are calculated for each decile. Finally, we compare the curve between the theoretical and sample quantile to the 45^0 degree line. The curve closest to this line indicates the best fitting distribution. From our QQ-plots, we argue that the distribution of THCC follows a Negative Binomial distribution (see Figure 3).
Finally, we consider two different regressions

- LLD is a parametric model:

$$\log(THCC) = \alpha_0 + \alpha_1 \log(HHINC) + \alpha_2 log^2(HHINC) + \sum \gamma Factors + \epsilon, \tag{3}$$

- GAMNB is a GAM model with the distribution of THCC belonging to the Negative Binomial family and with the following relationship between expected response and explanatory

$$log(\mathbb{E}(THCC)) = \beta_0 + s(HHINC) + \sum \theta Factors, \tag{4}$$

- where *Factors* include URBAN, HSIZE, ETHNIC, WA, EDUCH, GENDER, PRO

The coefficients of these models for the three years are presented in Table 4. Figure 4 is the boxplot of ASPE samples. The ASPE samples for the three years in the boxplot and the Kruskal-Wallis test show that the mean ASPE sample of the GAMNB model is significantly smaller than the mean of the ASPE of the LLD model. We conclude that the model GAMNB fits the data better than the LLD model. We now analyze in detail the results for the model GAMNB in 2006, 2009 and 2011.

Comparing the effect of HHINC in the two models for these three years, the HHINC coefficients in model LLD are not significant while for model GAMNB, the smooth functions are significantly different from zero. Moreover, Figure 5 describes the smooth function of total household calorie intake on household income suggesting a rather convex curve in the center of the income distribution for these three years. This shows that increasing household income leads to an increase of calorie intake at low levels of income, then at given high levels of income, the number of calorie tends to be stable as income increase. For very high income, there are different trends, either stability in 2006 and 2009 or a continuously increasing pattern in 2011. However, we see that the confidence intervals for very high income is quite large.

The coefficients of the variable URBAN are negative and significant (except in 2011) which shows that households in urban areas consume significantly less calories than those in the rural areas. This makes sense for at least three reasons. First, households in rural areas tend to consume a higher percentage of rich calories foods such as rice and staple foods. In contrast, the diets of urban households are more diversified with higher percentages of fruits, meats, fish and drink. Second, although household incomes in urban sites are higher than in rural sites, the price of food, and consequently the price of calories in rural areas are much lower than in urban areas. Lastly, the higher proportion of manual work in rural sites results in people needing more calorie intake.

Household size coefficients are all positive and significant for the three years. In addition, the value of the coefficient increases with the number of household members. The results are normal since we estimate the total household calories. Larger families lead to consume more calorie.

The coefficient for the variable ETHNIC representing the HAN nationality reveals a different behavior through years. The coefficients are insignificant for the three years.

The coefficient for other family characteristics such that WA are significant

except in 2011. The factor does not have a direct impact on household calorie intake but it depends on other household characteristics such as income or location.
The highest attained education level of the head of family reveals a different behavior according to the year and it is difficult to predict a general impact of education on the level of calorie consumption.
The coefficients corresponding to the gender of the head of the household is significantly negative for the three years. It means that households with a male head consume less calorie than those with a female head.
The PRO variable coefficients are very different for each year. There are several coefficients which are significant with positive and negative values while there are some provinces that do not have impact on per capita intake. It is very complicated to find an explanation since the eating behavior also depend economic characteristics as well as traditional culture of each region.

4.3 Income calorie elasticities

In this section, we focus on estimating the income calorie elasticities. For the two models LLD and GAMNB, the formula for elasticities are respective

- LLD model

$$\frac{\partial \log(THCC)}{\partial \log(HHINC)} = \alpha_1 + 2\alpha_2 \log(HHINC) \tag{5}$$

- GAMNB model

$$\frac{\partial log(\mathbb{E}(THCC))}{\partial \log(HHINC)} = s'(HHINC) \times (HHINC) \tag{6}$$

The average value of elasticities for the two models for the three years are summarized in Table 2. All the values are generally small (ranging from 0.07 to 0.11) but comparable with the elasticities in the paper of Nie and Sousa-Poza (2016).

Table 2: Income calorie elasticity for LLD model and GAM model

Year	2006	2009	2011
LLD	0.0780	0.1077	0.1245
GAMNB	0.0701	0.0962	0.1098

5 Conclusion

This paper has presented a comprehensive analysis of the relationship between calorie intake with other economic characteristics of households in China using the Chinese Health and nutrition survey in 2006, 2009 and 2011. The data set is analyzed with some semiparametric models as well as the traditional parametric model. By applying the cross validation procedure and simulation results, we have argued that semiparametric models involving a distribution for the response which belongs to the Negative Binomial distribution family outperform the traditional log-log Gaussian specification. Results for semiparametric models indicate a positive and significant effect of household income on per capita intake in China which is found by many previous authors for the same database.
The smooth curves for the three years suggest a behavior of Chinese households with regard to food demand. For very low income, the total calorie increases with income. Then, at a given medium level of income, Chinese households tend to maintain the number of calorie. Finally, at very high income, the trend varies, either stabilize, decrease and increase.
The calorie-income calorie elasticities are positive but small for all Chinese household which however demonstrates the efficiencies of income-mediated policies focused at fighting against food insecurity in China.

Acknowledgment

This research was funded as part of TAASE project, GloFoodS meta program, INRA - CIRAD, France. In addition, the authors would like to express their deepest thanks to Thibault LAURENT for his cooperation on the programming.

References

Bhargava, A. (2008). *Food, economics, and health*. Oxford:: Oxford University Press.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, Volume 38. SIAM.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 89–102.

Gibson, J. and S. Rozelle (2002). How elastic is calorie demand? parametric, nonparametric, and semiparametric results for urban papua new guinea. *Journal of Development Studies 38*, 23–46.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models*, Volume 43. CRC Press.

Marra, G. and S. N. Wood (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis 55(7)*, 2372–2387.

Nie, P. and A. Sousa-Poza (2016). A fresh look at calorie-income elasticities in china. *China Agricultural Economic Review 8*(1).

Ogundari, K. and A. Abdulai (2013). Examining the heterogeneity in calorie–income elasticities: A meta-analysis. *Food Policy 40*, 119–128.

Popkin, B. (2003). The nutrition transition in the developing world. *Development Policy Review 21*(5-6), 581–597.

Racine, J. and C. Parmeter (2014). *Data-Driven Model Evaluation: A Test for Revealed Performance, in Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics,*. Oxford University Press.

Ruppert, D., M. Wand, and R. L. Carroll (2004). *Semiparametric Regression*. Cambridge Univ Press.

Tian, X. and X. Yu (2015). Using semiparametric models to study nutrition improvement and dietary change with different indices: The case of china. *Food Policy 53*, 67–81.

Vu, H. (2009). Analysis of calorie and micronutrient consumption in vietnam. *Development and Policies Research Center Working Paper Series* (2009/14).

Wood, S. (2006). *Generalized Additive Models: An introduction with R*. Chapman and Hall/CRC.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 95–114.

Wood, S. N. and N. H. Augustin (2002). Gams with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological modelling 157(2)*, 157–177.

Xiang, D. and G. Wahba (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica 6*, 675–692.

Zucchini, W. (2000). An introduction to model selection. *Journal of mathematical psychology 44*(1), 41–61.

Table 3: Description CHNS data 2006, 2009 and 2011

Variable	2006 (3612 obs)	2009 (3785 obs)	2011 (3572 obs)
THCC	5392.77 (2359.16)	5405.69 (2421.71)	4957.12 (2324.49)
HHINC	24607.27 (20060.7)	34829.4 (27172.19)	40392.34 (30507.06)
RURAL	67.64 %	66.87 %	66.41 %
URBAN	32.36 %	33.13 %	33.59 %
HSIZE2	32.17 %	35.56 %	38.16 %
HSIZE3	27.74 %	26.66 %	25.28 %
HSIZE4	20.99 %	19.02 %	17.53 %
HSIZE5	19.1 %	18.76 %	19.04 %
Han0	12.35 %	12.89 %	12.46 %
Han	87.65 %	87.11 %	87.54 %
FEMALE	83.08 %	81.59 %	82.11 %
MALE	16.92 %	18.41 %	17.89 %
WA0	11.13 %	9.01 %	7.67 %
WA1	88.87 %	90.99 %	92.33 %
EDUCH1	42.97 %	41.53 %	40.85 %
EDUCH2	44.52 %	46.61 %	45.16 %
EDUCH3	12.51 %	11.86 %	14 %
Liaoning	11.3 %	11.2 %	11.31 %
Heilongjiang	11.57 %	11.76 %	11.51 %
Jiangsu	10.71 %	10.78 %	11.28 %
Shandong	11.3 %	11.04 %	11.48 %
Henan	10.71 %	10.99 %	9.8 %
Hubei	10.16 %	10.54 %	10.78 %
Hunan	11.54 %	10.57 %	11.51 %
Guangxi	10.88 %	11.97 %	11.17 %
Guizhou	11.82 %	11.15 %	11.17 %

THCC and HHINC are the mean, the other is the percentage for each level.

Figure 2: Density of THCC

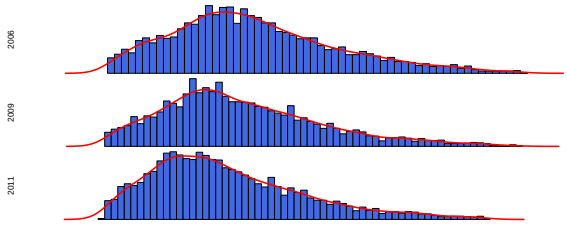


Figure 3: Fitting QQ-plot with Negative Binomial

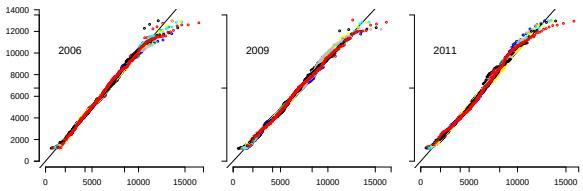
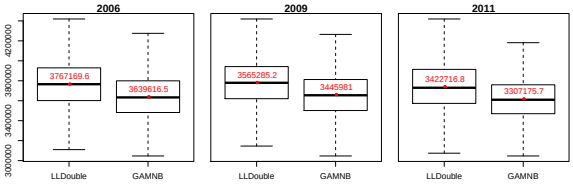


Figure 4: Boxplot of revealed performance test in 2006, 2009 and 2011



The Kruskal-Wallis test between the 2 models in each year have p-value $< 2.2e-16$.

Figure 5: The smooth term s(HHINC) in 2006, 2009 and 2011

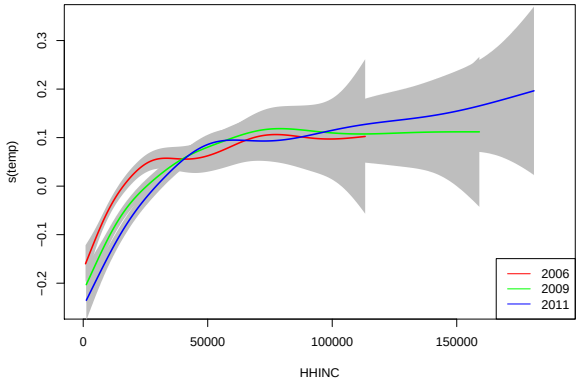


Table 4: Coefficient table for LLD and GAMNB in 2006, 2009 and 2011

	2006		2009		2011	
	LLdouble	GAMNB	LLdouble	GAMNB	LLdouble	GAMNB
(Intercept)	6.018 ***	8.187 ***	6.656 ***	8.165 ***	6.642 ***	8.039 ***
	(0.525)	(0.033)	(0.512)	(0.03)	(0.566)	(0.032)
log(HHINC)	0.355 **		0.182 .		0.13	
	(0.111)		(0.104)		(0.114)	
log(HHINC)^2)	-0.014 *		-0.004		0	
	(0.006)		(0.005)		(0.006)	
s(HHINC)		***		***		***
URBAN	-0.058 ***	-0.051 ***	-0.061 ***	-0.055 ***	-0.023	-0.025 .
	(0.015)	(0.014)	(0.014)	(0.013)	(0.015)	(0.015)
HSIZE3	0.32 ***	0.321 ***	0.312 ***	0.31 ***	0.287 ***	0.291 ***
	(0.017)	(0.016)	(0.016)	(0.015)	(0.017)	(0.016)
HSIZE4	0.472 ***	0.479 ***	0.43 ***	0.438 ***	0.415 ***	0.434 ***
	(0.019)	(0.018)	(0.018)	(0.017)	(0.019)	(0.019)
HSIZE5	0.634 ***	0.641 ***	0.646 ***	0.656 ***	0.613 ***	0.631 ***
	(0.02)	(0.019)	(0.018)	(0.018)	(0.02)	(0.019)
HAN	0.04 .	0.037 .	0.026	0.026	0.039 .	0.032
	(0.022)	(0.021)	(0.021)	(0.02)	(0.022)	(0.021)
WA1	0.038 .	0.031	0.087 ***	0.096 ***	0.055 *	0.03
	(0.022)	(0.021)	(0.022)	(0.021)	(0.025)	(0.024)
EDUCH2	0.016	0.013	-0.012	-0.014	0.041 **	0.036 *
	(0.015)	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)
EDUCH3	-0.013	-0.019	-0.071 **	-0.079 ***	0.006	0.003
	(0.024)	(0.023)	(0.022)	(0.021)	(0.023)	(0.022)
MALE	-0.155 ***	-0.125 ***	-0.16 ***	-0.132 ***	-0.141 ***	-0.116 ***
	(0.018)	(0.017)	(0.016)	(0.015)	(0.017)	(0.017)
Heilongjiang	0.03	0.027	0.033	0.022	0.091 ***	0.099 ***
	(0.027)	(0.026)	(0.025)	(0.024)	(0.027)	(0.026)
Jiangsu	0.116 ***	0.11 ***	0.119 ***	0.121 ***	0.148 ***	0.147 ***
	(0.028)	(0.026)	(0.026)	(0.025)	(0.028)	(0.027)
Shandong	0.069 *	0.065 *	0.032	0.037	0.157 ***	0.157 ***
	(0.027)	(0.026)	(0.026)	(0.025)	(0.027)	(0.026)
Henan	-0.01	-0.01	0.059 *	0.057 *	0.14 ***	0.129 ***
	(0.028)	(0.027)	(0.026)	(0.025)	(0.029)	(0.028)
Hubei	0.076 **	0.069 *	0.075 **	0.081 **	0.186 ***	0.205 ***
	(0.028)	(0.027)	(0.026)	(0.025)	(0.028)	(0.027)
Hunan	0.043	0.035	0.018	0.034	0.121 ***	0.103 ***
	(0.027)	(0.026)	(0.026)	(0.025)	(0.027)	(0.026)
Guangxi	-0.052 .	-0.039	0.07 **	0.062 *	0.179 ***	0.17 ***
	(0.028)	(0.026)	(0.026)	(0.025)	(0.028)	(0.027)
Guizhou	0.026	0.031	0.007	0.007	0.008	0.01
	(0.028)	(0.027)	(0.026)	(0.025)	(0.028)	(0.027)