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

Huong TRINH THI, Michel SIMIONI, Christine THOMAS-AGNAN
Toulouse School of Economics*

May 4, 2016

Abstract

The recent research on calorie intake and income relationship abounds with parametric models which give 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, supported by simulation data, shows that GAM models outperform the parametric models. Using data from CHNS in 2006, 2009 and 2011, we argue a positive and statistically significant relationship between household calorie intake and household income for the poor, then the impact of increasing income on calorie consume will slow down for the middle class and the rich. In addition, we find that income-calorie elasticities are general small, ranging from 0.07 to 0.12.

Keywords: Calorie intake and income, generalized additive model, exponential family, CHNS data, revealed performance test, cross validation criterion.

1 Introduction

Food consumption patterns of population is a critical problem for both developed and developing countries and its impacts are summarized in

*Contact author : trinhthihuong@vcu.edu.vn

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 the developing country and has been concerned with the issue of the burden of nutrition-related non communicable diseases (NR-NCDs), a follow stage after receding famine. China is not an exception with the large ratio of people having NR-NCDs, and cancer is the first cause of death in urban China and the second one in rural China (Zhao et al. (2010)).

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 study seems to suggest that specifying the relationship between the response variable and the explanatory variables plays an important role. The relationship is potentially nonlinear ((Deaton, 1997; Popkin, 2003, see, for instance)), however, among 99 income-calorie elasticities collected in the paper, 86 values are estimated by using parametric model with logarithm or square in order to capture non-linearity. Only few papers use semiparametric models to deal with the issue of non linearity (see, for instance Gibson and Rozelle, 2002; Vu et al., 2009; Nie and Sousa-Poza, 2016; Tian and Yu, 2015). In the case of China, as summarized in Nie and Sousa-Poza (2016), current researches appear to validate the view that elasticities vary substantially, even among studies using the same dataset. For example, in the recently 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 method 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. On our knowledge, there are three main arguments that can be advanced to support to 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 et al. (2014). The approach of the test comprises building the distribution function of a model's *true error* (Efron, 1982), then, the Kruskal and Wallis (1952) test is applied to compare whether the *expected* true error is statistically smaller for one model than another. The simulation data on our paper provides strong evidence that the revealed performance test choose the best fitting model in the case pre-known the data generating process (DGP).

The paper is organized as follow. Section 2 discusses semiparametric regression using penalized spline smoothing. Section 3 gives the idea of cross - validation criterion 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 traditional approach regression. The conclusion and several suggestions for policy maker 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 $\mathbb{E}(Y)$. They 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 form

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

where b, a and c are arbitrary function, 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 solving the model, we advocate 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. Comparing between parametric models and semi-parametric models, on logical grounds, there is no compelling reason to argue that semi-parametric model will perform better than parametric model. To go further, our discussion will point to a cross validation criterion, named a test for revealed performance, initiated by Racine et al. (2014). The principal 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 models on the remaining validation data, next we compute average square prediction error (ASPE) (we know the Y values for the evaluation data, hence this delivers 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

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

A closer look at the ASPE value indicates that one model outperform the other but depending on particular division of the data in two independent

subsets. To overcome the limitation, we put forward the claim that we need to repeat the process many times, say $S = 1000$ time. Thus, the two sample ASPEs generated by the procedure are used to discriminate between the two modes. 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. 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* in Efron (1982), and also the test does not require the same type of models.

Although there has been simulation on the Racine and Parameter's work, further evidence supporting the cross validation test on our problem, say the income calorie intake relationship, needs to be considered. In the next step, we conduct three simulation exercises where each model includes response variable y and explanatory continuous variables x and factor fac. The simulation models are summarized as follow

- (LLdouble), parametric model

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

- (GauNL), y follows Gaussian distribution

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

- (NBNL), y follows Negative Binomial distribution

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

Where, $s(x) = 1 + 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10}$. The factor fac has 2 levels 0 and 1 and the number observations in each time simulation is $n = 3000$. By simulation the data LLdouble, we would like to assess a data having linear relationship in the term log-log square, on the other side, the function $s(x)$ is complicated enough to describe the non-linearity between the covariate and the response variable. In addition, the two data GauNL and NBNL will test how the criterion react with the difference choice of the distribution in the semi-parametric model.

For what follows, we fit each simulation with models LLdouble, GauNL and NBNL, then apply the cross validation criterion with $n_1 = 2500$, $n_2 =$

500 and $S = 1000$. The boxplot of ASPE sample is reported in the Figure 1. Also, the Table 1 show the Kruskal-Wallis test to compare the mean of the two samples ASPE between the pair of fitting. Visually, for the LLdouble simulated data, the mean of the three samples are quite equal. Along similar lines, for the GauNL and NBNL simulation, the mean of the sample are well discriminated between parametric model and semi-parametric model. Kruskal-Wallis test shows small $p-$value for the pair LLdouble-GauNL and LLdouble-NBNL in the three simulations, indicating significant difference in the mean. However, between the pair GauNL and NBNL, the $p-$value is large. Thus, we can not reject the hypothesis that the mean of the two samples are significantly different. We would like to conclude that in the case of simulation data, there is no difference between applying parametric model and semiparametric model when the data has linear relationship. However, in the case non-linear relationship, the cross validation test will choose for the right model (here is GauNL data and NBNL data with fitting GauNL and NBNL model).

Figure 1: Boxplot of the ASPE sample from simulation data

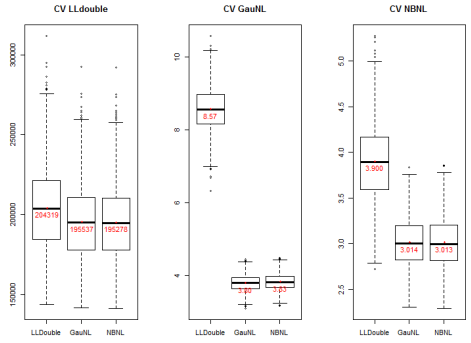


Table 1: $p-$ value of Kruskal-Wallis test for simulation data

	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 depend on each year since many families that migrate from one community to a new one are not followed anymore. Here, we focus on the year 2006, 2009 and 2011 with the purpose of finding the 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 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 the significant relationship, Nie and Sousa-Poza (2016) concludes that Chinese household is quite successful in maintaining the amount of calorie as income vary. After filtering observations, our data set is summarized in Table 3.

These data include calorie intake (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 equal sharing calorie in 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, nonretirement earnings, and other, deflated in 2006. There is a rapid increase of income in China. The average household income is 24607 Yuan, 34829 Yuan and 40392 Yuan in 2006,

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

2009 and 2011 respectively.

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

4.2 Results of model fitting

We apply the traditional parametric model in the issue, say log-log double model (LLD), and semiparametric model with generalized additive models, say GAM, to the Chinese data with several specifications. For the implement in R for GAM model, we use the package mgASPE proposed by S.Wood.

The choice of the distribution for THCC will impact the quality of fitting model. From the histogram of THCC for each year (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 interested distribution, the sample quantile and theoretical quantile are calculated for each decile. Finally, we compare the plot between the theoretical-sample quantile and the 45^0 degree line. The better fitting between the point and the line indicates the sample THCC following the corresponding distribution. From the result of fitting QQ-plot, we argue that the distribution of THCC follows Negative Binomial distribution (see Figure 3).

Finally, we consider two different regressions

- (LLD) The parametric model:

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

- (GAMNB) GAM model and the distribution of THCC belongs to Negative Binomial:

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

- Where factors include URBAN, HSIZE, ETHNIC, WA, EDUCH, GENDER, PRO

The coefficients of these models in the three years are presented in Table 4.

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

Comparing the coefficient of HHINC in the 2 models in these three years, HHINC coefficients in model LLD are not significant while on models 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 convex curve in the center of the income distribution in 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, the number shows difference trends, either stably in 2006 and 2009 or increases continuously in 2011. However, we see that the confidence interval for very high income is quite large thus the trends will vary.

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 in 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. While in 2006 it is positive and significant, it changes to insignificant in 2009 and 2011.

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 consume.

The gender of the head of the household is significantly negative in the 3 years. It means that households with the male heading member consume less calorie than those with female.

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 the reason since their behavior of eating also depend on their economic problems as well as their traditional culture.

4.3 Income calorie elasticities

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

- 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 in the three years are summarized in table 2. All the number are generally small (range from 0.07 to 0.11) but compared with the elasticities in the paper of Nie and Sousa-Poza (2016).

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

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

5 Conclusion

This paper has presented a comprehensive analysis of calorie intake with other economical characteristic for households in China using the Chinese Health and nutrition survey in 2006, 2009 and 2011. The data set is analyzed with semiparametric models as well as the traditional parametric model. By applying the cross validation criterion and simulation results, we have argued that semiparametric models involving a distribution for the response which belongs to Negative Binomial distribution outperform the traditional log-log Gaussian specification. Results from semiparametric models indicate a positive and significant effect of household income on per capita intake in China which is found by many previous authors in the same database.

The smooth curves in the three years suggest a behavior of Chinese households on food demanding. For very low income, a income increases, the total calorie also increases. And then, at a given 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 and small for all Chinese household which demonstrates the efficiencies of income-mediated policies focused at fighting against food insecurity in China.

References

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

Deaton, A. (1997). *The analysis of household serveys.* The Johns Hopkins 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. J. and R. J. Tibshirani (1990). *Generalized additive models*, Volume 43. CRC Press.

Kruskal, W. H. and W. A. Wallis (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association 47*(260), 583–621.

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

Racine, J., L. Su, and A. Ullah (2014). *The Oxford 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. L. et al. (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.

Zhao, P., M. Dai, W. Chen, and N. Li (2010). Cancer trends in china. *Japanese journal of clinical oncology 40*(4), 281–285.

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 %
HSIZE1	5.32 %	6.5 %	6.63 %
HSIZE2	26.85 %	29.06 %	31.52 %
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 %
EL0	0.33 %	0.26 %	0.95 %
EL1	99.67 %	99.74 %	99.05 %
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 %
EDUCH0	22.98 %	21.22 %	20.27 %
EDUCH1	19.99 %	20.32 %	20.58 %
EDUCH2	29.93 %	33.84 %	32.75 %
EDUCH3	14.59 %	12.76 %	12.4 %
EDUCH4	6.81 %	6.37 %	6.24 %
EDUCH5	5.7 %	5.5 %	7.75 %
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 2006,2009 and 2011

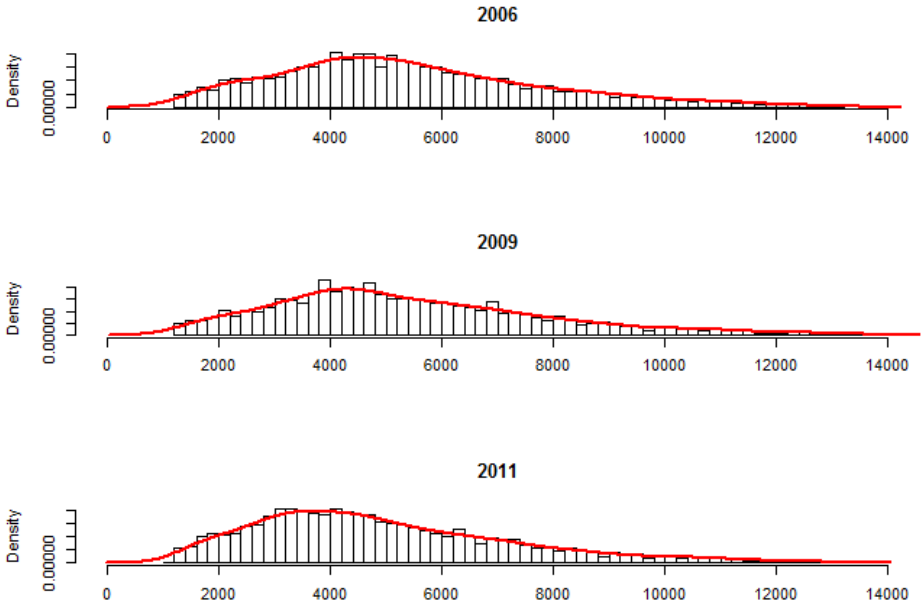


Figure 3: Fitting QQ-plot with Negative Binomial in 2006, 2009 and 2011

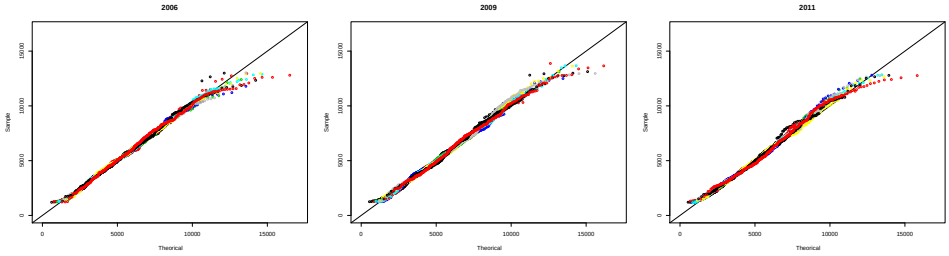
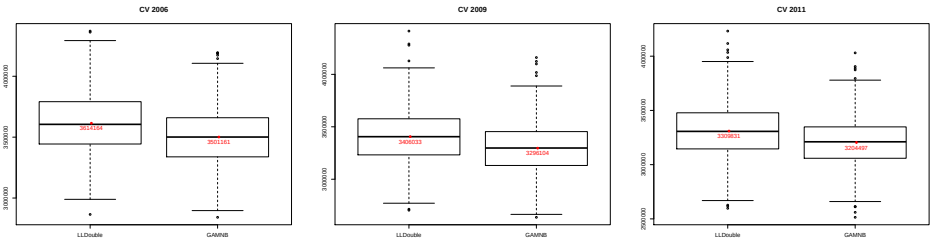


Figure 4: Boxplot of Cross validation criterion in 2006, 2009 and 2011



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

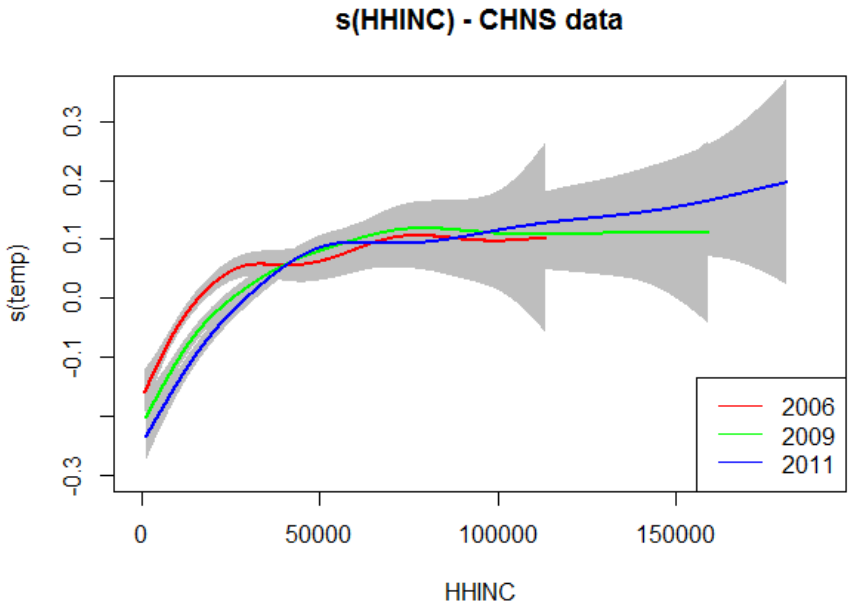


Table 4: Coefficient table for LLD and GAMNB in 2006, 2009 and 2011

	2006		2009		2011	
	LLdouble	GAMNB	LLdouble	GAMNB	LLdouble	GAMNB
(Intercept)	6.43 ***	7.619 ***	7.254 ***	7.62 ***	7.041 ***	7.574 ***
	(0.515)	(0.041)	(0.485)	(0.037)	(0.545)	(0.041)
log(HHINC)	0.18 .		-0.015		-0.013	
	(0.106)		(0.099)		(0.11)	
$log(HHINC)^2$	-0.006		0.005		0.006	
	(0.006)		(0.005)		(0.006)	
URBAN1 0.01	-0.049 ***	-0.044 **	-0.051 ***	-0.048 ***	-0.01	-0.013
(0.013)	(0.015)	(0.014)	(0.013)	(0.013)	(0.015)	(0.014)
HSIZE2	0.566 ***	0.598 ***	0.545 ***	0.576 ***	0.471 ***	0.499 ***
	(0.03)	(0.028)	(0.026)	(0.025)	(0.028)	(0.027)
HSIZE3	0.814 ***	0.857 ***	0.779 ***	0.819 ***	0.694 ***	0.733 ***
	(0.031)	(0.029)	(0.027)	(0.026)	(0.029)	(0.028)
HSIZE4	0.963 ***	1.012 ***	0.895 ***	0.944 ***	0.818 ***	0.872 ***
	(0.031)	(0.03)	(0.028)	(0.027)	(0.03)	(0.029)
HSIZE5	1.123 ***	1.172 ***	1.115 ***	1.166 ***	1.018 ***	1.072 ***
	(0.032)	(0.03)	(0.028)	(0.027)	(0.03)	(0.029)
ETHNIC1	0.046 *	0.045 *	0.025	0.027	0.043 *	0.033
	(0.021)	(0.02)	(0.02)	(0.019)	(0.021)	(0.021)
WA1	0.048 *	0.04 *	0.094 ***	0.099 ***	0.054 *	0.03
	(0.021)	(0.02)	(0.02)	(0.019)	(0.024)	(0.023)
EDUCH1	0.02	0.02	0.044 *	0.044 *	0.033 .	0.029
	(0.019)	(0.018)	(0.018)	(0.017)	(0.02)	(0.019)
EDUCH2	0.031 .	0.026	0.005	0.004	0.058 **	0.052 **
	(0.018)	(0.017)	(0.017)	(0.016)	(0.018)	(0.018)
EDUCH3	-0.004	-0.003	-0.003	-0.006	0.023	0.014
	(0.022)	(0.021)	(0.021)	(0.02)	(0.023)	(0.023)
EDUCH4	-0.036	-0.041	-0.05 .	-0.057 *	0.031	0.024
	(0.029)	(0.027)	(0.028)	(0.026)	(0.03)	(0.029)
EDUCH5	0.02	0.02	-0.06 *	-0.065 *	0.009	0.008
	(0.032)	(0.03)	(0.03)	(0.029)	(0.029)	(0.028)
GENDER1	-0.078 ***	-0.062 ***	-0.068 ***	-0.051 ***	-0.063 ***	-0.047 **
	(0.017)	(0.017)	(0.016)	(0.015)	(0.017)	(0.017)
Heilongjiang	0.018	0.015	0.043 .	0.03	0.1 ***	0.106 ***
	(0.026)	(0.024)	(0.024)	(0.023)	(0.026)	(0.025)
Jiangsu	0.129 ***	0.129 ***	0.138 ***	0.139 ***	0.156 ***	0.157 ***
	(0.026)	(0.025)	(0.025)	(0.024)	(0.027)	(0.026)
Shandong	0.068 **	0.064 *	0.043 .	0.045 .	0.156 ***	0.156 ***
	(0.026)	(0.025)	(0.025)	(0.024)	(0.026)	(0.025)
Henan	-0.015	-0.014	0.063 *	0.054 *	0.132 ***	0.126 ***
	(0.027)	(0.026)	(0.025)	(0.024)	(0.028)	(0.027)
Hubei	0.079 **	0.075 **	0.084 ***	0.087 ***	0.195 ***	0.212 ***
	(0.027)	(0.026)	(0.025)	(0.024)	(0.027)	(0.026)
Hunan	0.044 .	0.038	0.006	0.021	0.115 ***	0.098 ***
	(0.026)	(0.025)	(0.025)	(0.023)	(0.026)	(0.025)
Guangxi	-0.048 .	-0.035	0.073 **	0.063 **	0.182 ***	0.173 ***
	(0.026)	(0.025)	(0.024)	(0.023)	(0.027)	(0.026)
Guizhou	0.036	0.046 .	0.016	0.018	0.006	0.01
	(0.027)	(0.026)	(0.025)	(0.024)	(0.027)	(0.026)