Journal of Statistical and Econometric Methods, vol.1, no.1, 2012, 1-12
ISSN: 2241-0384 (print), 2241-0376 (online)
International Scientific Press, 2012

Estimation of Parameters in Weighted Generalized Beta Distributions of the Second Kind

Yuan Ye[1], Broderick O. Oluyede[2] and Mavis Pararai[3]

Abstract

This paper applies the class of weighted generalized beta distribution of the second kind (WGB2) as descriptive models for size distribution of income. The properties of WGB2 including mean, variance, coefficient of variation(CV), coefficient of skewness(CS), coefficient of kurtosis(CK) are presented. Other properties including top-sensitive index, bottom-sensitive index, mean logarithmic deviation(MLD) index and Theil index obtained from generalized entropy(GE) are applied in this paper. WGB2 proved to be in the generalized beta-F family of distributions, and maximum likelihood estimation(MLE) is used to obtain the parameter estimates. WGB2 is fitted to U.S. family income (2001-2009) data with different values of the parameters. The empirical results show the length-biased distribution provides the best relative fit.

Mathematics Subject Classification : 62F10
Keywords: WGB2, Income distribution, Entropy, Beta-F family, MLE

1 Introduction

Generalized beta distribution of the second kind (GB2) has been widely used in income distribution. It provides a good description of income distri-

[1] Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA 30460, e-mail: yy00053@georgiasouthern.edu
[2] Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA 30460, e-mail: boluyede@georgiasouthern.edu
[3] Department of Mathematics, Indiana University of Pennsylvania, Indiana, PA 15705, e-mail: pararaim@iup.edu

Article Info: *Received* : November 29, 2011. *Revised* : December 26, 2011
Published online : February 28, 2012

bution and captures the characteristics of size distribution of income such as: skewness, has a peak in low-middle range, and long right hand tail [7]. McDonald [7] adopted different distributions as models for size distribution of income, and found that GB2 provides the best relative fit.

Weighted distribution provides an approach to dealing with model specification and data interpretation problems. Fisher [3] and Rao [12] introduced and unified the concept of weighted distribution. Cox [1] and Zelen [17] used it to present length biased sampling. The usefulness and applications of weighted distribution to biased samples in various areas including medicine, ecology, reliability, and branching processes can also be seen in Nanda and Jain [9], Gupta and Keating [2], Oluyede [10] and in references therein.

Suppose Y is a non-negative random variable with its natural pdf $f(y;\theta)$, θ is a parameter, then the pdf of the weighted random variable Y^w is given by:

$$f^w(y;\theta,\beta) = \frac{w(y,\beta)f(y;\theta)}{\omega}, \tag{1}$$

where the weight function $w(y,\beta)$ is a non-negative function, that may depend on the parameter β, and $0 < \omega = E(w(Y,\beta)) < \infty$ is a normalizing constant.

This paper applies the class of WGB2 as descriptive models for the size distribution of income. The properties of WGB2 such as moments and generalized entropy (GE) are presented in Section 2. Section 3 applies WGB2 in the size distribution of income. WGB2 is fitted to U.S. family income data (2001-2009) with different values of the parameter k in Section 4. The parameter estimates of WGB2 are obtained and empirical results are presented. Section 5 contains an application of the results to US family nominal income for the years 2001 to 2009.

2 The Distribution and Special Cases

Weighted generalized beta distribution of the second kind (WGB2) with polynomial weight function $w(y) = y^k$ is a very flexible five-parameter distribution. The probability density function (pdf) of WGB2 is given by:

$$g_{WGB2}(y;a,b,p,q,k) = \frac{ay^{ap+k-1}}{b^{ap+k}B(p+\frac{k}{a},q-\frac{k}{a})[1+(\frac{y}{b})^a]^{p+q}}, \tag{2}$$

where $y > 0$, $a, b, p, q > 0$ and $-ap < k < aq$.

WGB2 includes GB2 as a special case, it also includes several other weighted distributions as special or limiting cases: WGG (weighted generalized gamma), WB2 (weighted beta of the second kind), WSM (weighted Singh Maddala),

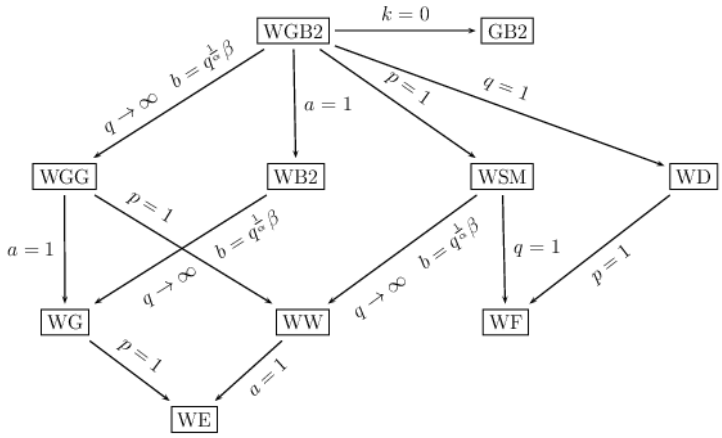

Figure 1: Graph Tree

WD (weighted Dagum), WG (weighted gamma), WW (weighted Weibull) and WE (weighted exponential). The distribution tree is given below:

Jones and Faddy [5] introduced the concept of generalized beta-F distribution:

$$g_F(y;\alpha,\beta) = B(\alpha,\beta)^{-1} f(y)[F(y)]^{\alpha-1}[1-F(y)]^{\beta-1},$$

where $f(y)$ is the derivative of $F(y)$, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function. Sepanski and Kong [14] applied the generalized beta-F family in size distribution of income and obtained the parameter estimates. They concluded that log-F followed by GB2 provides the best relative fit.

3 Related Results

From our previous work, we obtain the moments of WGB2 with weight function $w(y) = y^k$, that is

$$E_{G_{WGB2}}(Y^j) = \frac{b^j B(p+\frac{k}{a}+\frac{j}{a}, q-\frac{k}{a}-\frac{j}{a})}{B(p+\frac{k}{a}, q-\frac{k}{a})}. \tag{3}$$

The corresponding mean and variance are given by

$$\mu_{G_{WGB2}} = E_{G_{WGB2}}(Y) = \frac{bB(p+\frac{k}{a}+\frac{1}{a}, q-\frac{k}{a}-\frac{1}{a})}{B(p+\frac{k}{a}, q-\frac{k}{a})}, \tag{4}$$

and

$$Var_{G_{WGB2}}(Y) = b^2\left[\frac{B(p+\frac{k+2}{a}, q-\frac{k+2}{a})}{B(p+\frac{k}{a}, q-\frac{k}{a})} - (\frac{p+\frac{k+1}{a}, q-\frac{k+1}{a}}{p+\frac{k}{a}, q-\frac{k}{a}})^2\right]. \quad (5)$$

respectively. The coefficient of variation (CV) is given by

$$CV_{WGB2} = \sqrt{\frac{B(p+\frac{k+2}{a}, q-\frac{k+2}{a})B(p+\frac{k}{a}, q-\frac{k}{a})}{B^2(p+\frac{k+1}{a}, q-\frac{k+1}{a})} - 1}. \quad (6)$$

Similarly, the coefficient of skewness and coefficient of kurtosis are

$$CS = \frac{E[Y^3] - 3\mu E[Y^2] + 2\mu^3}{\sigma^3}, \quad (7)$$

and

$$CK = \frac{E[Y^3] - 4\mu E[Y^3] + 6\mu^2 E[Y^2] - 3\mu^4}{\sigma^4}, \quad (8)$$

where

$$\mu = \mu_{G_{WGB2}}, \quad \sigma = \sqrt{Var_{G_{WGB2}}(Y)}, \quad E[Y^2] = \frac{b^2 B(p+\frac{k}{a}+\frac{2}{a}, q-\frac{k}{a}-\frac{2}{a})}{B(p+\frac{k}{a}, q-\frac{k}{a})},$$

$$E[Y^3] = \frac{b^3 B(p+\frac{k}{a}+\frac{3}{a}, q-\frac{k}{a}-\frac{3}{a})}{B(p+\frac{k}{a}, q-\frac{k}{a})}, \quad and \quad E[Y^4] = \frac{b^4 B(p+\frac{k}{a}+\frac{4}{a}, q-\frac{k}{a}-\frac{4}{a})}{B(p+\frac{k}{a}, q-\frac{k}{a})}.$$

The generalized entropy (GE) is widely used to measure inequality trends and differences. It is primarily used in income distribution. Generalized entropy $I(\alpha)$ for WGB2 is given by:

$$I(\alpha) = \frac{B(p+\frac{k}{a}+\frac{\alpha}{a}, q-\frac{k}{a}-\frac{\alpha}{a})B^{-\alpha}(p+\frac{k}{a}+\frac{1}{a}, q-\frac{k}{a}-\frac{1}{a}) - B^{1-\alpha}(p+\frac{k}{a}, q-\frac{k}{a})}{\alpha(\alpha-1)B^{1-\alpha}(p+\frac{k}{a}, q-\frac{k}{a})}, \quad (9)$$

where $\alpha \neq 0$ and $\alpha \neq 1$. The bottom-sensitive index is $I(-1)$, and the top-sensitive index is $I(2)$. Moreover, the mean logarithmic deviation (MLD) index and Theil index are:

$$I(0) = log\frac{B(p+\frac{k+1}{a}, q-\frac{k+1}{a})}{B(p+\frac{k}{a}, q-\frac{k}{a})} - \frac{\psi(p+\frac{k}{a})}{a} - \frac{\psi(q-\frac{k}{a})}{a}, \quad (10)$$

and

$$I(1) = \frac{\psi(p+\frac{k+1}{a})}{a} - \frac{\psi(q-\frac{k+1}{a})}{a} - log\frac{B(p+\frac{k+1}{a}, q-\frac{k+1}{a})}{B(p+\frac{k}{a}, q-\frac{k}{a})}. \quad (11)$$

respectively.

4 Estimation of Parameters

For WGB2 with weight function $w(y) = y^k$, the pdf can be written as:

$$g_w(y; a, b, p, q) = B^{-1}\left(p + \frac{k}{a}, q - \frac{k}{a}\right)\left(\frac{a}{b}\right)\left[\left(\frac{y}{b}\right)^a\right]^{p+\frac{k}{a}}\left[1 + \left(\frac{y}{b}\right)^a\right]^{-(p+q)}.$$

If we set $F(y) = 1 - (1 + (\frac{y}{b})^a)^{-1}$, then $f(y) = \frac{a}{b}(\frac{y}{b})^{a-1}[1 + (\frac{y}{b})^a]^{-2}$ and

$$g_w(y; a, b, p, q) = B^{-1}\left(p + \frac{k}{a}, q - \frac{k}{a}\right) f(y)[F(y)]^{p+\frac{k}{a}-1}[1 - F(y)]^{q-\frac{k}{a}-1}. \quad (12)$$

Clearly, this distribution belongs to the $beta - F$ class of distributions with

$$F(y) = 1 - \left[1 + \left(\frac{y}{b}\right)^a\right]^{-1} = \frac{y^a}{b^a + y^a}. \quad (13)$$

Let $\theta = (a, b, p, q)^T$ be a column vector of parameters associated with the income distribution. The income distribution is given by:

$$P_i(\theta) = B^{-1}\left(p + \frac{k}{a}, q - \frac{k}{a}\right)\int_{F(y_{i-1})}^{F(y_i)} t^{p+\frac{k}{a}-1}(1-t)^{q-\frac{k}{a}-1}dt, \quad (14)$$

where $P_i(\theta)$ denotes the estimated proportion of the population in the i^{th} interval of the r income groups defined by the interval $I_i = [y_{i-1}, y_i]$. The multinomial likelihood function for the data is given by:

$$N!\prod_{i=1}^{r}\frac{[P_i(\theta)]^{n_i}}{n_i!},$$

where $n_i, i = 1, 2, ..., r$ denotes the observed frequency in the i^{th} group and $N = \sum_{i=1}^{r} n_i$. We maximize:

$$L(\theta) = \sum_{i=1}^{r} n_i ln P_i(\theta),$$

where $P_i(\theta) = \int_{F(y_{i-1})}^{F(y_i)} h(t)dt$, and $h(t) = B^{-1}(p + \frac{k}{a}, q - \frac{k}{a})t^{p+\frac{k}{a}-1}(1-t)^{q-\frac{k}{a}-1}$. Sepanski and Kong [14] pointed out that obtaining P_i by computing the cdf of a beta random at $F(y_{i-1})$ and $F(y_i)$ can reduce the complexity of programming required to calculate the integrations.

The first derivative with respect to $\theta = (a, b, p, q)^T$ are:

$$\frac{dL(\theta)}{d\theta} = \sum_{i=1}^{r}\frac{n_i}{P_i(\theta)} \cdot \frac{dP_i(\theta)}{d\theta}. \quad (15)$$

The partial derivative equations of $P_i(\theta)$ with respect to a, b, p, q are given by:

$$\frac{\partial P_i(\theta)}{\partial a} = h(F(y_i))\frac{b^a y_i^a (lny_i - lnb)}{(b^a + y_i^a)^2} - h(F(y_{i-1}))\frac{b^a y_{i-1}^a (lny_{i-1} - lnb)}{(b^a + y_{i-1}^a)^2} + \int_{F(y_{i-1})}^{F(y_i)} \frac{kh(t)}{a^2}\left[\Psi\left(p + \frac{k}{a}\right) - \Psi\left(q - \frac{k}{a}\right) + ln\frac{1-t}{t}\right]dt, \quad (16)$$

$$\frac{\partial P_i(\theta)}{\partial b} = -ab^{a-1}\left[\frac{h(F(y_i))y_i^a}{(b^a + y_i^a)^2} - \frac{h(F(y_{i-1}))y_{i-1}^a}{(b^a + y_{i-1}^a)^2}\right], \quad (17)$$

$$\frac{\partial P_i(\theta)}{\partial p} = \int_{F(y_{i-1})}^{F(y_i)} h(t)\left[-\Psi\left(p + \frac{k}{a}\right) + \Psi(p+q) + lnt\right]dt, \quad (18)$$

and

$$\frac{\partial P_i(\theta)}{\partial q} = \int_{F(y_{i-1})}^{F(y_i)} h(t)\left[-\Psi\left(q - \frac{k}{a}\right) + \Psi(p+q) + ln(1-t)\right]dt, \quad (19)$$

where $\Psi(x) = \frac{d}{dx}[\Gamma(x)]$. Using the equations (15)-(18) in equation (14), we can obtain the gradient functions of $L(\theta)$ with respect to parameters a, b, p, q.

The partial derivative equation (15) exists when $k > 0$. If $k = 0$, the partial derivative equation of P_i with respect to a is given by:

$$\frac{\partial P_i(\theta)}{\partial a} = b^a\left[\frac{h(F(y_i))(y_i)^a(ln(y_i) - ln(b))}{(b^a + y_i^a)^2} - \frac{h(F(y_{i-1}))(y_{i-1})^a(ln(y_{i-1}) - ln(b))}{(b^a + y_{i-1}^a)^2}\right]. \quad (20)$$

5 Applications

In this section, we obtain parameter estimates based on our previous discussions and results. WGB2 was fitted to U.S. family nominal income for 2001-2009[4]. The groups consist of families whose income are in the corresponding income interval $I_i = [y_{i-1}, y_i)$, the n_i/N are the observed relative frequencies ($N = \sum n_i$).

The common way to obtain parameter estimates is to maximize the multinomial likelihood function. Since the likelihood function is nonlinear and complicated, we use MATLAB to search for the maximum value of multinomial

likelihood function.[5] The results of this estimation for 2001, 2005, and 2009 are reported in Tables 2-4.

Based on the sum of squares error (sse) value we can conclude that: the length-biased WGB2 ($k = 1$) provides a better fit than GB2 ($k = 0$) and other WGB2 ($k = 2, 3, 4$). If we plug the estimated parameters in the partial derivative equations in Section 4, we obtain the values of these partial derivative equations in Table 5-7. From the tables, we find that these values are close to zero or very small, this means that the estimated parameters that we obtained are precise and effective.

Since we have already obtained estimates of parameters for WGB2 in Section 4, and found that the length-biased WGB2 provides the best fit to income distribution, we can apply these estimated parameters from length-biased WGB2 model to obtain the estimates of the mean, variance, coefficient of variation, skewness and kurtosis, bottom sensitive index, top-sensitive index, MLD index and Theil index. The results are presented in Table 8.

6 Concluding Remarks

In this paper, the weighted generalized beta distribution of the second kind (WGB2) was fitted to U.S. family income data (2001-2009). The maximum likelihood estimation (MLE) is used for estimating the parameters of the income distribution model. The results showed that the length-biased WGB2 provides the best relative fit to income data. Based on previously obtained descriptive measures for WGB2, we estimate the mean, variance, coefficient of variation, coefficient of skewness, coefficient of kurtosis, bottom-sensitive index, top-sensitive index, MLD index and Theil index of the income data.

ACKNOWLEDGEMENTS. The authors wish to express their gratitude to the referees and editor for their valuable comments.

[4]The data were taken from the Census Population Report

[5]By setting different initial values and using 'fminsearchbnd' to search for the maximum log likelihood values

References

[1] D. R. Cox, *Renewal Theory*, Barnes & Noble, New York, 1962.

[2] R.C. Gupta and J.P. Keating, Relations for Reliability Measures Under Length Biased Sampling, *Scandinavian Journal of Statistics*, **13**(1), (1985), 49 - 56.

[3] R.A. Fisher, The Effects of Methods of Ascertainment upon the Estimation of Frequencies, *Annals of Human Genetics*, **6**(1), (1934), 439 - 444.

[4] S.P. Jenkins, Inequality and GB2 Income Distribution, *ECINQE, Society for study of Economic Inequality, Working Paper Series*, (2007).

[5] M.C. Jones and M.J. Faddy, A Skew Extension of the t distribution, with Applications, *Journal of the Royal Statistical Society. Series B*, **65**(1), (2003), 159 - 174.

[6] C. Kleiber and S. Kotz, Statistical Size Distributions in Economics and Actuarial Sciences, Wiley, New York, 2003.

[7] J.B. McDonald, Some Generalized Functions for the Size Distribution of Income, *Econometrica*, **52**(3), (1984), 647-663.

[8] J.B. McDonald and Y.J. Xu, A Generalization of the Beta Distribution with Application, *Journal of Econometrics*, **69**(2), (1995), 133 - 152.

[9] K.A. Nanda and K. Jain, Some Weighted Distribution Results on Univariate and Bivariate Cases, *Journal of Statistical Planning and Inference*, **77**(2), (1999), 169 - 180.

[10] B.O. Oluyede, On Inequalities and Selection of Experiments for Length-Biased Distributions, *Probability in the Engineering and Informational Sciences*, **13**(2), (1999), 129 - 145.

[11] G.P. Patil and R. Rao, Weighted Distributions and Size-Biased Sampling with Applications to Wildlife and Human Families, *Biometrics*, **34**(6), (1978), 179 - 189.

[12] C.R. Rao, On Discrete Distributions Arising out of Methods of Ascertainment, *The Indian Journal of Statistics*, **27**(2), (1965), 320 - 332.

[13] A. Renyi, On Measures of Entropy and Information, *In Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, **1**(1960), University of California Press, Berkeley, (1961), 547 - 561.

[14] J.H. Sepanski and L. Kong, A Family of Generalized Beta Distribution For Income, *International Business* , **1**(10), (2007), 129 - 145.

[15] C.E. Shannon, A Mathematical Theory of Communication, *The Bell System Technical Journal*, **27**(7), (1948), 379 - 423.

[16] C.E. Shannon, A Mathematical Theory of Communication, *The Bell System Technical Journal*, **27**(10), (1948), 623 - 656.

[17] M. Zelen, Problems in Cell Kinetics and Early Detection of Disease, *Reliability and Biometry*, **56**(3), (1974), pp.701 - 726.

Table 1: U.S. family nominal income for 2001-2009

$[y_{i-1}, y_i)$	observed relative frequencies n_i/N								
(thousand)	2001	2002	2003	2004	2005	2006	2007	2008	2009
[0,15)	13	13.4	12.9	12.6	13	13.3	13.2	12.9	12.4
[15,25)	11.9	12	11.3	11.2	11.5	11.6	11.6	11.4	11.4
[25,35)	11.1	11	10.5	11.1	10.8	11	10.9	10.6	10.5
[35,50)	14.1	14.1	14	14.1	14.2	14.1	14	14.5	14.8
[50,75)	18.1	17.6	18	18.2	18.1	18.1	17.7	18	17.9
[75,100)	11.5	11.9	12	11.6	12.1	12	12.2	12.5	12.6
[100,150)	11.9	11.9	12.7	12.5	12	11.9	12.3	12.3	12.2
[150,200)	4.4	4.3	4.7	4.7	4.3	4.4	4.4	4.2	4.3
[200,∞)	3.8	3.7	4	4	4	3.6	3.7	3.7	3.9

Table 2: Estimated parameters of WGB2 for income distribution (2001)

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
a	1.403	1.405	0.669	0.4487	0.3376
b	16.66	16.488	11408.457	5463.111	15213.5
p	0.999	0.288	0.000001	0.001681	0.037333
q	3.963	4.629	470.178	154.668	187.036
sse*10000	1.208234	1.208159	2.509558	8.463283	13.220825

Table 3: Estimated parameters of WGB2 for income distribution (2005)

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
a	1.423	1.383	0.646	0.432	0.339
b	15.502	16.621	12800.159	6595.86	4379.966
p	0.961	0.271989	0.000003	0.004804	0.00976
q	3.582	4.648	446.0185	156.662	127.893
sse*10000	0.520372	0.518495	2.109663	8.049098	13.146924

Table 4: Estimated parameters of WGB2 for income distribution (2009)

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
a	1.090	1.092	0.654	0.440	0.341
b	26.596	26.695	2324.550	3550.805	3964.261
p	1.413	0.493	0.000005	0.001079	0.002585
q	7.209	8.148	157.039	125.837	124.988
sse*10000	0.471333	0.467639	1.078739	6.085883	10.444168

Table 5: Values of partial derivative equations of WGB2 (2009)

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
$\frac{\partial L(\theta)}{\partial a}$	0.0255	-0.0029	0.0397	-0.1732	-0.1347
$\frac{\partial L(\theta)}{\partial b}$	-0.0017	-0.0001	0	0	0
$\frac{\partial L(\theta)}{\partial p}$	-0.0105	-0.0048	-0.0791	-0.1077	-0.0625
$\frac{\partial L(\theta)}{\partial q}$	0.007	-0.0005	0	0.0004	0

Table 6: Values of partial derivative equations of WGB2 for (2005)

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
$\frac{\partial L(\theta)}{\partial a}$	-0.0069	-0.0341	0.0028	-0.0922	0.0033
$\frac{\partial L(\theta)}{\partial b}$	-0.0021	-0.0017	0	0	0
$\frac{\partial L(\theta)}{\partial p}$	0.0118	0.0422	-0.1075	-0.097	-0.0638
$\frac{\partial L(\theta)}{\partial q}$	-0.0007	-0.0058	0.0002	0.001	-0.0001

Table 7: Values of partial derivative equations of WGB2 (2001)

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
$\frac{\partial L(\theta)}{\partial a}$	-0.0283	-0.0502	-0.2614	-0.6607	3.2358
$\frac{\partial L(\theta)}{\partial b}$	-0.0031	-0.0123	0	0	0
$\frac{\partial L(\theta)}{\partial p}$	0.0173	-0.1533	-0.1243	-0.1381	-0.0055
$\frac{\partial L(\theta)}{\partial q}$	-0.0009	0.0327	0.0003	0.0011	-0.0028

Table 8: Estimated statistics for income distribution ($k=1$ in WGB2)

Year	Est.mean	Est.Var	Est.CV	Est.CS	Est.CK
2001	6.759468	38.091976	0.913070	-4.242920	24.013806
2005	6.719345	39.169450	0.931423	-4.114463	25.864490
2009	6.629841	37.111483	0.918864	-4.195311	16.203708
Year	Est.I(-1)	Est.I(2)	Est.MLE	Est.Theil	
2001	1.150014	0.416849	0.396929	1.758663	
2005	1.252677	0.433774	0.410256	1.789568	
2009	1.033116	0.422155	0.402408	3.567591	