A STUDY ON THE INFLUENCING FACTORS OF CPI BIAS IN DIFFERENT GROUPS IN CANADA

by

Haoyu Wang

Bachelor of Finance, Jilin University, 2015

Master of Economics, Shandong University of Finance and Economic, 2018

A Report Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the Graduate Academic Unit of Economics

Supervisor: Herb Emery, PhD, Dept. of Economics

Examining Board: Yuri V. Yevdokimov, PhD, Dept. of Economics, Chair

Philip Leonard, PhD, Dept. of Economics

Ted Mcdonald, PhD, Dept. of Economics

This report is accepted by the Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

June, 2019

©Haoyu Wang, 2019

Abstract

The official CPI may mismeasure the cost of living for different groups of people. This paper investigates the factors that might explain the size of the CPI bias measured by Emery and Guo (2019) for the years 1999 to 2015. We apply the partial least squares method (PLS) estimates of CPI bias for 10 provinces and 14 sub-groups to determine which subcomponents of the CPI are influential on the CPI bias. The result shows that gasoline, fuel and clothing are important factors affecting the CPI bias of each group. However, when we group the samples in 2010 as the cutoff point, this effect is more significant in the later stage, while the influence factors of each group in the early stage have a large inter-group difference and the exchange rate has an important impact on the CPI bias of each group, especially in the later stage.

Acknowledgement

I would like to thank my supervisor Dr. Emery, for his expert guidance, suggestions, supervision and support. I would also like to thank the other members of my committee, for their very useful comments as readers.

I also appreciate my family members and friends for their support. Without their strong support, I could not complete my MA report.

Table of Contents

1. Introduction

Emery and Guo (2019) used an Engel Curve Approach to estimate the size of the the Consumer Price Index bias for 10 provinces and 14 subgroups for the years 1999 to 2015 following the work of Costa (2001), Hamilton (2001a), Hamilton (2001b) and Brzozowski (2006). The CPI has a number of potential sources of bias. Across identifiable groups like seniors and non-seniors, consumption patterns may differ from those of average or representative households used to construct the CPI hence the CPI may mismeasure cost of living and changes in cost of living of specific sub-groups of the population. An Engel Curve approach to estimating the size of the CPI bias exploits Engel's Law which states that the share of food in a household's budget is inversely related to the household's real income. Changes in a household's food budget share can be used to infer changes in the purchasing power of a household's income real income. Changes in the food budget share that are inconsistent with changes in real income measured as nominal income deflated by the official CPI can be used as a measure of the CPI bias.

What is the reason the official CPI misrepresents the cost of living for subgroups? One possibility is that weights used to construct the official CPI reflect consumption budget shares of various goods and services of a representative household consumption patterns and preferences. The preferences and consumption patterns of various commodities of different subgroup may differ. Therefore, when constructing the real CPI, different

groups should assign different weights to the same commodity according to different consumption weights. The official CPI gives the same weight to the consumption of goods in different groups, which results in the bias of CPI, which will result in the inefficiency of social welfare security policies. Therefore, exploring the influencing factors of CPI bias in different groups, so as to construct real and effective CPI for different groups, can promote the effectiveness and pertinence of the formulation of social security and tax policies and improve the overall social utility. For this possibility, the most direct way to test the influencing factors of CPI bias of different groups is to calculate the differences of contribution degrees of different commodity price changes to the group.

To determine what factors might explain the sizes of Emery and Guo's estimates of the CPI bias different groups, I apply the PLS method to determine which subcomponents of the CPI may be influential for household cost of living not captured by the official CPI. PLS is a multivariate statistical data analysis method which uses component extraction technology and comprehensive screening approach to process the data in the regression system to realize the selection and importance ordering of variables in the modeling process. Using the PLS method to explore the relationship between subcomponents of the CPI and CPI bias in each group, we find that gasoline, fuel and clothing are important factors contributing to CPI bias in each group, and this effect becomes significant after 2010. Before 2010, the influencing factors of CPI bias showed

significant differences in different groups.

2. Determining influence with Partial Least Squares methods

In this paper, partial least squares regression method is used to construct the influencing factor model of CPI bias among different populations. PLS is applicable to both single dependent variable and multiple dependent variables. This empirical study only involves the single dependent variable of CPI bias, so only PLS modeling of single dependent variable is elaborated.

In classical least-squares regression, if two or more explanatory variables are highly correlated, it is not easy to distinguish their individual influences on the explained variables (Greene, 2003). We try to take the change of the price index of different commodities as the explanatory variable, and the price of these commodities is affected by the overall social income level, monetary policy, consumption structure characteristics and other factors, which will inevitably lead to certain multicollinearity among the variables. If multicollinearity is found, the following treatment can be used. 1) If we don't care about specific regression coefficients, but only about the ability of the entire equation to predict the explained variable, we can usually ignore multicollinearity (assuming that the entire equation is significant). 2) If we care about a specific regression coefficient, but multicollinearity does not affect the significance of the variable we care about, we can ignore it. 3) If multicollinearity affects the significance of the variables concerned, we need to increase the sample size, eliminate the variables

leading to severe collinearity, or modify the model settings (Chen, 2014). We are concerned about the significance of CPI in each category, so collinearity problem cannot be ignored. In this paper, partial least squares method (PLS) is selected to solve this problem.

PLS is a multivariate statistical data analysis method proposed by Wold and Albano (1983), which has unique advantages in solving the problem of multiple correlation of independent variables. It is different from the traditional regression method. This method does not directly consider overall regression modeling of the dependent variable and independent variable in the regression modeling, but uses component extraction technology and comprehensive screening approach to process the data in the regression system and get some new comprehensive variables which have the best explanation and the biggest influence on the the dependent variable. The established model identifies the information and noise in the system and can effectively overcome the adverse effects caused by multiple correlations among independent variables in regression modeling (Wold, 2001). In addition, PLS has a whole set of auxiliary analysis technology, which can realize the selection of variables and the evaluation of the model in the process of modeling.

PLS has been applied in several scientific fields. Hulland (1999) reviewed the literature on PLS in the field of strategic management, and pointed that models such as PLS can help strategic management researchers to achieve new insights. Pistonesi et al. (2006)

apply PLS to a chemistry experiment. When a third component was present (phenol), the resolution of the new mixture was impossible because the resorcinol signal was overlapped to the phenol signal. PLS method allowed the resolution of the three components in the mixtures and can deal with the experimental data, thus to establish a simple, rapid and sensitive method for simultaneous determination of hydroquinone, resorcinol and phenol in samples. Krishnan (2011) pointed out that PLS methods are particularly suited to the analysis of relationships between measures of brain activity and of behavior or experimental design and illustrated the application of PLS method in neural imaging with numerical examples. Monk et al. (2013) applied this method to the field of geographic and biological sciences and studied the influence of a series of landscape variables on cold water refugia in rivers by PLS. His explanatory variable was a simulated shelter in geothermal infrared imagery collected from a stretch of the Cains river in New Brunswick, Canada, in late July 2008 and July 2009. Explanatory variables include subindexes of climate, stream morphology, geology, etc. This method successfully avoids collinearity of these indexes and the results suggest that median temperatures of tributary catchments are driven by their position within the landscape including slope in addition to the density of wetlands and mixed forest within the upstream catchment. These literature are characterized by a large number of independent variable indicators, one or more dependent variable indicators, and these independent variable indicators have strong collinearity. PLS method can be used to give the ranking of the influence degree of each indicator.

2.1. PLS modeling steps

Let the vector matrix constituted by dependent variable be Y, and the set constituted by the p independent variables incorporated into the model be $X = (X_1, X_2 \cdots X_p)$. In order to investigate the relationship between Y and X, let's say we have n observed samples (The samples selected in this paper are n, so we have n×1 dependent variable vector matrix $Y_{n\times1}$ and n×p independent variable matrix $X_{n\times p}$).

Step 1: Let the data vector of Y after normalization be $F_0 = (F_{01}, F_{02} \cdots F_{0q})$, since Y is a single dependent variable, q=1; The normalized matrix of X is $E_0 = (E_{01}, E_{02} \cdots E_{0p})_{n\times p}$ and $x_i^* = \frac{X_i - \overline{X_i}}{S_i}, i = 1,2 \cdots p$; S_i is the standard deviation of X_i, $y^* = \frac{Y - \overline{Y}}{S_Y}$, S_Y is the standard deviation of Y.

Step 2: Let t_1 is the first principal component of E_0, $t_1 = E_0 w_1$, W_1 is the first axis of E_0, and $\|w_1\| = 1$; Let u_1 be the first component of F_0, $u_1 = F_0 c_1$, c_1 is the first axis of F_0, and $\|c_1\| = 1$. Meanwhile, t_1 and u_1 are required to contain the variation information of E_0 and F_0 as much as possible, t_1 has the highest correlation with u_1, that is, t_1 not only synthesizes the information of X to the maximum extent, but also has the strongest explanatory power for Y, which can be expressed by the formula as follows:

$$\max(E_0 w_1, F_0 c_1), \text{st} \begin{cases} w_1 w_2 = 1 \\ c_1 c_2 = 1 \end{cases}$$

The regression equations of E_0 to t_1 and F_0 to t_1 and u_1 were solved respectively,

$E_0 = t_1 p_1^{'} + E_1$, $F_0 = u_1 q_1^{'} + F_1^*$, $F_0 = t_1 r_1^{'} + F_1$,

Where the regression coefficient vector is $p_1 = \frac{E_0^{'} t_1}{P t_1 P^2}$, $q_1 = \frac{F_0^{'} u_1}{P u_1 P^2}$, $r_1 = \frac{F_0^{'} t_1}{P t_1 P^2}$; E_1, F_1^* and F_1 are residual matrices of the above three regression equations respectively. Since t_1 is a linear combination of E_0, we can finally get the regression equation of Y to t_1. If the regression of Y to t_1 meets the accuracy requirement (which can be set by cross validation), proceed to the next step, otherwise, go back to the second step and conduct principal component extraction and regression analysis for the residual matrix E_1 and F_1 again.

Step 3: Assume that principal component extraction is carried out h(h>=1) times and the accuracy requirement of regression equation is satisfied. So now we have h components, $t_1, t_2 \cdots t_h$, and carry out the regressions of E_0 and F_0 on $t_1, t_2 \cdots t_h$, then we have $E_0 = t_1 p_1^{'} + t_2 p_2^{'} \cdots + t_h p_h^{'}$ and $F_0 = t_1 r_1^{'} + t_2 r_2^{'} \cdots + t_h r_h^{'}$, where $t_h = E_{h-1} w_h$, $p_h = \frac{E_{h-1}^{'} t_h}{P t_h P^2}$, $r_h = \frac{F_{h-1}^{'} t_h}{P t_h P^2}$. For that $w_h = E_{h-1}^{'} F_0 / P E_{h-1}^{'} F_0 P$, $E_{h-1} = t_h p_h^{'} + E_h$, and $t_1, t_2 \cdots t_h$ are all linear representations of E_0, therefore we have $F_0 = \sum_{m=1}^{h} r_m E_0 w_m^*$, where $w_m^* = \prod_{j=1}^{h-1} (I - w_j p_j^{'}) w_h$, and I is the identity matrix.

By $y^* = F_0$, $x_i^* = E_{0i}$, We can derive that $y^* = \alpha_1 x_1^* + \alpha_2 x_2^* + \cdots + \alpha_p x_p^*$, $\alpha_i = \sum_{m=1}^{h} r_m w_{mi}^*$, where w_{mi}^* is the *i*-th component of w_m^*.

Step 4: The regression equation of F_0 is mapped to the regression equation of Y to X by

inverse operation of normalization.

2.2. The determination of PLS component number

In many cases, partial least squares regression equation does not need to use all components for regression modeling. The question is how many ingredients to choose. The cross-validation method is widely used in academia to determine the optimal number of components. The general idea is to divide n sample observations into two parts. The first part is to first remove the sample i, and then use $t_1, t_2 \cdots t_h$ components fit into a regression equation to the n-1 samples left. The second part is to substitute the elimination sample i into the regression equation and get the fitted value. $\hat{y}_{h(-i)}$, for each sample point i (i =1,2... n) repeat the above steps. Thus, the sum of the squares of the prediction errors of Y is defined as $PRESS_h = \sum_{i=1}^{n} (y_i - \hat{y}_{h(-i)})^2$. In addition, take all the sample points and take the components $t_1, t_2 \cdots t_h$ fitting a regression equation. Let the fitting value of the i-th sample point be $\hat{y}_{hi}$, so we can get the sum of the squared errors as $SS_h = \sum_{i=1}^{n} (y_i - \hat{y}_{hi})^2$. Define $Q_h^2 = 1 - PRESS_h / SS_{h-1}$, when $Q_h^2 \geq 0.0975$ (SIMCA-P14.1 default), it is beneficial to introduce h-th component.

2.3. PLS assisted analysis technology

2.3.1. Discovery of singular values

The existence of sample singular value will disturb the regression, and then reduce the prediction accuracy of the model. Partial least squares regression determines the

existence of singular value by calculating the contribution rate of each sample point to the extracted h components. Let the cumulative contribution rate of the i-th sample point to the component $t_1, t_2 \cdots t_h$ be $T_i^2 = \frac{1}{n-1}\sum_{m-1}^{h}\frac{t_{mi}^2}{s_m^2}$. In the statistical sense, T_i^2 should not be too large, otherwise the analysis will be biased. So construct the Tracy statistic $\frac{n^2(n-h)}{h(n^2-1)}T_i^2 \sim F(h, n-h)$. When $T_i^2 \geq \frac{h(n^2-1)}{n^2(n-h)}F_{0.05}(h, n-h)$, it can be considered that at the significance level of 5%, sample point i contributes too much to component $t_1, t_2 \cdots t_h$, so that sample point i is considered as a singular value. When h=2, the discriminant condition becomes

$(\frac{t_{1i}^2}{s_1^2} + \frac{t_{2i}^2}{s_2^2}) \geq \frac{2(n-1)(n^2-1)}{n^2(n-2)}F_{0.05}(2, n-2)$. This is an ellipse, so we can make this ellipse in the t_1/t_2 plane. If all sample points fall within the ellipse, the singular value is considered to be nonexistent. If a sample point falls outside the ellipse, singular values are considered to exist.

2.3.2. Model evaluation

The scatter plot can clearly and intuitively analyze the relationship between independent variable and dependent variable in the univariate regression. This method is not feasible in multiple regression analysis because it is very difficult to directly observe the hyperplane formed by multidimensional data. And because there are collinearity problems between variables, it is not feasible to simply separate variables for analysis.

But the PLS algorithm's t_1-u_1 planar graph makes this possible. In PLS algorithm, the correlation between the principal components t_1 and u_1 of X and Y can reflect the correlation between the variable set X and the dependent variable Y. Plot the plane figure of t_1-u_1, with the abscissa of t_1 and the ordinate of u_1, and draw a scatter diagram composed of observation samples of the first principal component pair (t_1,u_1). If all samples can be arranged in a straight line in the figure, it indicates that there is a strong correlation between X and Y, and the model design will be reasonable.

2.3.3. The goodness of fit of the model

The goodness of fit of the model can be verified by analyzing the fitted value of the model $y(i=1,2\cdots,n)$ and the actual observed value $\hat{y}_i(i=1,2\cdots,n)$. If all the points can be uniformly distributed around the diagonals of the graph, it means that there is little difference between the fitted value of the model and the actual value. At this point, the fitting effect of this equation is satisfactory.

2.3.4. Variable importance in projection

In partial least squares regression analysis, the explanatory power of the ith variable Xi on the dependent variable Y is characterized by variable importance in projection (VIP_i), where:

$$VIP_i = \sqrt{\frac{p}{Rd(Y;t_1,t_2\cdots t_h)}\sum_{m=1}^{h} Rd(Y;t_m)w_{mi}^2}$$

w_{mi} is the i-th component of axis w_m to characterize the marginal contribution of X_i to

component t_m. For any m=1,2... h, $\sum_{i=1}^{p} w_{mi}^2 = w_m^{'} w_m = 1$. $Rd(Y;t_1,t_2\cdots t_h)$ and $Rd(Y;t_m)$ respectively represent cumulative explanatory power of $t_1,t_2\cdots t_h$ to Y and t_m's explanatory power to Y. If t_m has a strong explanatory power for Y, and X_i plays a significant role in the construction of t_m, it can be deemed that X_i has a strong explanatory power for Y, and the corresponding VIP_i value will be large, and vice versa. In economic practice, it is generally believed that when $VIP_i \geq 1$, the explanatory ability of variables is very strong; when $0.5 \leq VIP_i < 1$, the explanatory ability of variables is strong; when $VIP_i < 0.5$, the explanatory ability of variables is considered to be weak.

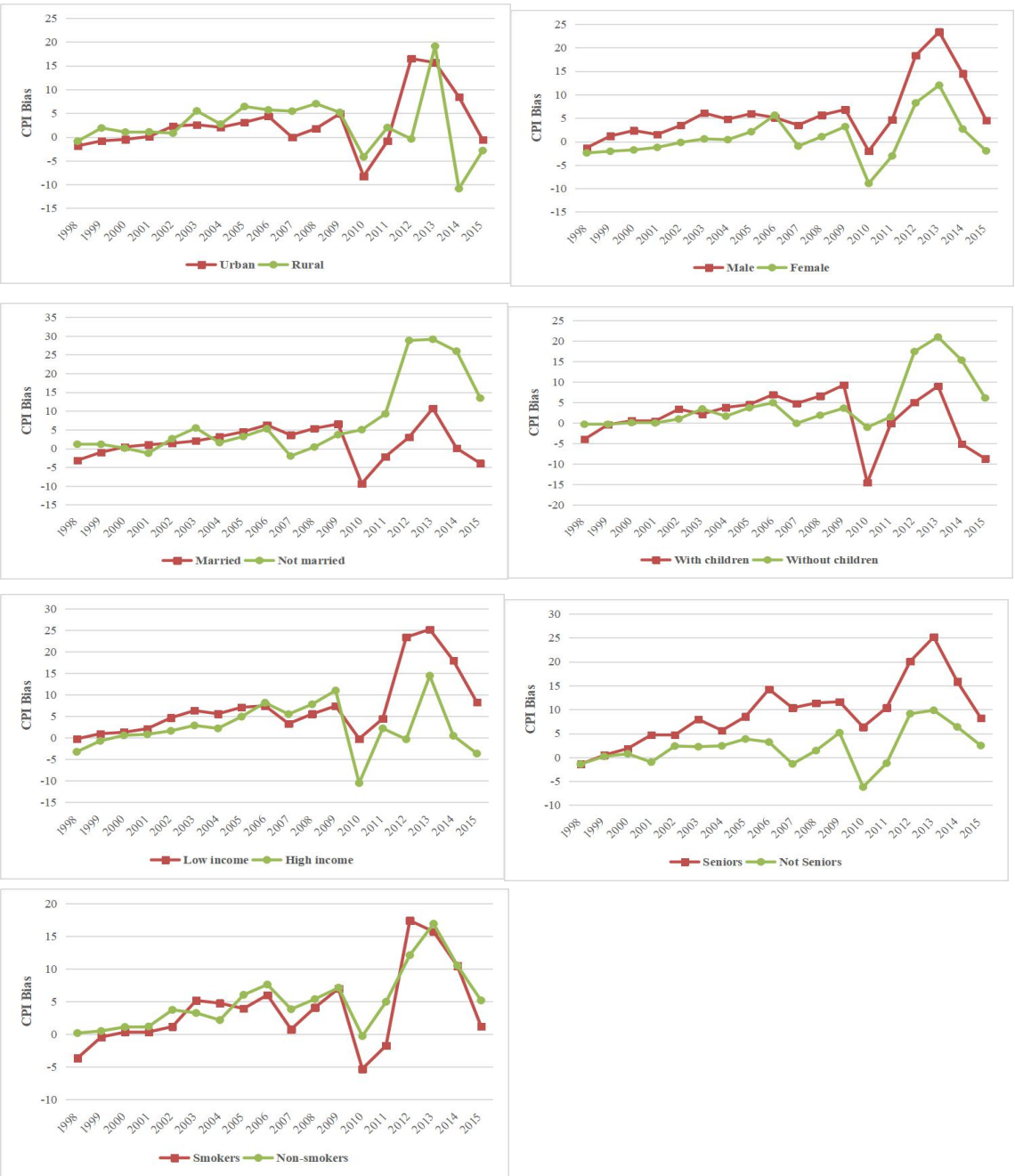
3. Data sources and descriptive statistics

In order to study the influencing factors of CPI bias estimated by Emery and Guo (2019) for different groups. I use bias estimates for seven pairs of groups for all 10 provinces for the years 1998 to 2015 as research objects: rural and urban, male and female, married and not married, with children and without children, low income and high income, seniors and not seniors, smokers and non-smokers. Such groupings are useful because they are relevant to the recipients of government welfare policies. See Emery and Guo (2019) for the detailed explanation for the estimation of the CPI bias.

Based on the data from Emery and Guo (2019), Figure 3.1 shows the time trend of the mean of bias in each province. It can be seen that the bias of most groups is positive, indicating that the official CPI generally underestimates the cost of living of each group. However, different groups were relatively stable before 2009, and then fluctuated greatly,

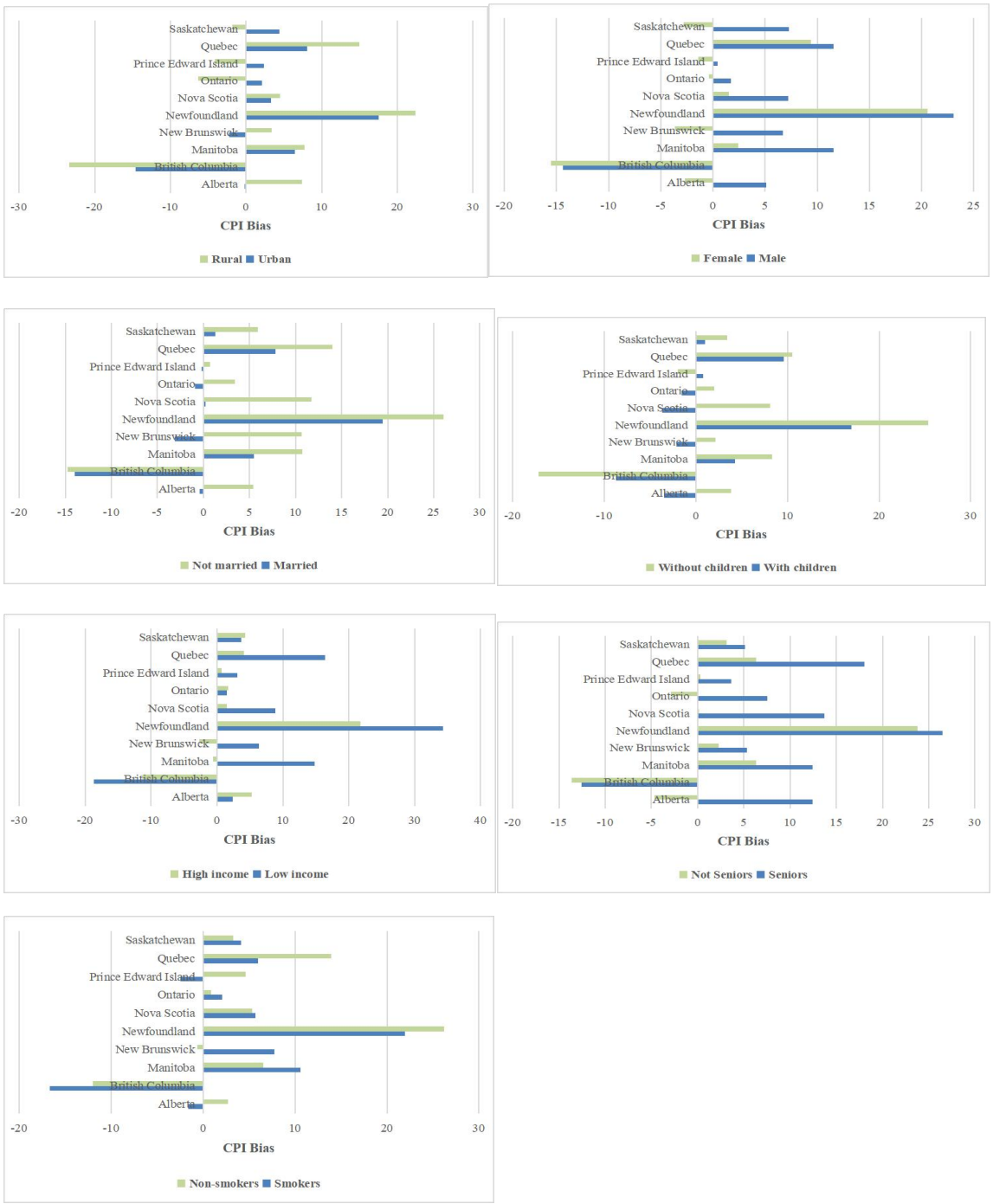
mainly reflecting the rise from 2010 to 2013 and the decline from 2013 to 2015. This may be related to the economic fluctuations and monetary policy after the financial crisis. Before 2009, the degree of bias among paired items of different groups was relatively similar. After 2010, the degree of bias among male was higher than that of female, the degree of bias among non-married people was higher than that of married people, the degree of bias among low-income people was higher than that of high-income people, and the degree of bias among the seniors was always higher than that of non-seniors. This suggests that since the financial crisis, the official CPI has underestimated the cost of living even more severely for male, non-married people, low-income people and seniors. Figure 3.2 shows the provincial differences of CPI bias mean from 1998 to 2015. It can be seen that the bias in British Columbia is generally negative, and the bias in other provinces is basically positive, while the bias in Newfoundland is the largest, followed by some groups in Quebec and Manitoba. Many groups in Prince Edward Island, Ontario and Nova Scotia provinces have different sign in term pairing, while most provinces have the same sign between groups.

Figure 3.1 Trends of CPI bias of subgroups, 1998-2015



Source: Emery and Guo (2019).

Figure 3.2 CPI bias of Canadian provinces



Source: Emery and Guo (2019).

In order to study what are the influencing factors of CPI bias in different groups, CPI sub component data of eight representative commodities are selected as independent variables, they are 1) Water, 2) Clothing and footwear (Clothing), 3) Natural gas, 4) Fuel oil and other fuels (Fuel), 5) Gasoline, 6) Rented accommodation (Rented), 7) Owned accommodation (Owned), 8) Electricity. These variables fully cover all aspects of the cost of living, with a certain representativeness. The data source is CANSIM database (Consumer Price Index, annual average, not seasonally adjusted. Table: 18-10-0005-01, formerly CANSIM 326-0021). The data are the annual data of 10 provinces from 1998 to 2015.

We have differential processing of all the variables to show the changes in the CPI of various goods. The advantage is that, on the one hand, we are more concerned about the impact of CPI changes on different groups of CPI bias, which can be reflected by difference form. On the other hand, the regression of difference data can eliminate the inter-provincial differences that do not change over time, making the results more accurate. Finally, we get the CPI change index of different commodities in each province from 1999 to 2015. Table 3.1 shows the correlation coefficients between the independent variables. It can be seen that the absolute values of the correlation coefficients among some independent variables exceed 0.3 or even reach 0.9, showing a strong correlation. Therefore, there is collinearity among independent variables and PLS is an appropriate choice for regression.

Table 3.1: Correlation coefficient between independent variables

Variable	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline
Rented	1							
Owned	0.3683	1						
Electricity	0.1566	-0.1033	1					
Water	0.3329	-0.0001	0.4224	1				
Natural gas	-0.2364	0.0633	-0.3486	-0.1898	1			
Fuel	-0.1264	0.2216	-0.1742	-0.0698	0.3183	1		
Clothing	-0.0299	-0.2641	0.0306	0.1126	-0.033	-0.2957	1	
Gasoline	-0.1596	0.176	-0.2127	-0.1396	0.2113	0.9423	-0.2694	1

Source: CANSIM database (Consumer Price Index, annual average, not seasonally adjusted. Table: 18-10-0005-01, formerly CANSIM 326-0021).

In order to explore the difference of this impact of the subcomponents of CPI on CPI bias before and after the crisis, we also divided the samples into two sub-samples from 1999 to 2009 and 2011 to 2015, taking 2010 as the cut-off point. Table 2 shows descriptive statistics of relevant variables. It can be seen that the bias of the non-married group fluctuates the most, 0.72, the low-income group and the seniors reach 0.5, while the other groups basically range from 0.2 to 0.3. And the late mean is generally higher than the early. The consumer price index for clothing and footwear has fallen, but for other items has risen. Among them, the increase rate of water and fuel oil and other fuels is large, exceeding 6, and gasoline exceeded 4. The increase rate of water and electricity in later stage is higher than that in early stage, while fuel oil and other fuels and gasoline

are larger in early stage. In the early stage, the growth rate of natural gas is fast, exceeding 4, while in the later stage, the change is negative. The price increase of owned accommodation in the early stage is higher than the rent price, while in the later period the increase is similar.

Table 3.2: Descriptive statistics of CPI bias changes and CPI component changes

Variable	1999-2015					1999-2009					2011-2015				
	Obs	Mean	St.D	Min	Max	Obs	Mean	St.D	Min	Max	Obs	Mean	St.D	Min	Max
Urban	170	0.07	10.45	-32.36	42.13	110	0.61	5.93	-15.51	16.02	50	1.53	15.08	-32.3	42.1
Rural	170	-0.12	23.53	-206.3	132.1	110	0.55	7.78	-19.72	21.37	50	0.27	41.26	-206.	132.
Male	170	0.34	10.90	-40.37	29.13	110	0.74	6.51	-13.86	15.99	50	1.31	15.90	-40.3	29.1
Female	170	0.03	11.42	-39.72	43.10	110	0.51	6.72	-20.06	18.85	50	1.39	16.84	-36.7	43.1
Married	170	-0.05	8.97	-34.50	20.97	110	0.88	4.88	-13.18	13.90	50	1.08	12.32	-24.5	20.9
Not married	170	0.72	13.60	-47.00	61.34	110	0.23	8.59	-23.84	23.32	50	1.68	20.76	-47.0	61.3
With children	170	-0.28	13.99	-63.26	42.61	110	1.20	5.19	-11.25	14.69	50	1.17	21.40	-63.2	42.6
Without	170	0.38	10.39	-32.23	53.20	110	0.36	6.49	-26.28	18.64	50	1.43	15.55	-32.2	53.2
Low income	170	0.50	11.70	-40.83	52.89	110	0.70	7.24	-27.94	20.63	50	1.69	17.02	-34.1	52.8
High income	170	-0.02	14.07	-48.73	71.94	110	1.30	6.16	-15.20	16.41	50	1.37	21.78	-48.7	71.9
Seniors	170	0.57	9.95	-30.62	26.30	110	1.19	7.65	-25.48	21.39	50	0.38	13.49	-30.0	26.3
Not Seniors	170	0.23	10.25	-32.50	32.13	110	0.60	6.99	-25.94	16.90	50	1.74	14.02	-32.5	32.1
Smokers	170	0.29	12.72	-49.62	60.22	110	0.97	6.65	-18.54	19.34	50	1.30	19.42	-42.5	60.2
Non-smokers	170	0.29	8.72	-24.88	22.70	110	0.63	6.79	-20.52	19.08	50	1.09	11.48	-24.8	22.7
Rented	170	1.59	1.20	-0.9	7.4	110	1.46	1.22	0.3	7.4	50	1.79	1.01	0.1	4.2
Owned	170	2.71	2.87	-2.9	19	110	3.23	3.21	-2.9	19	50	2.06	1.67	-2.2	6.1
Electricity	170	3.16	5.84	-18.5	30.6	110	2.91	5.18	-14.5	23.3	50	3.85	7.00	-18.5	30.6
Water	170	6.00	5.51	-5.2	25.2	110	4.89	5.22	-5.2	25.2	50	8.20	5.31	-1.9	21.5
Natural gas	102	2.62	18.64	-72.7	87.4	66	4.68	20.72	-72.7	87.4	30	-0.96	13.90	-35.2	30.6
Fuel	153	6.91	28.63	-75.8	63.3	99	7.14	28.36	-75.8	63.3	45	2.83	30.91	-60.1	50.7
Clothing	170	-0.04	1.62	-5.0	5.2	110	-0.31	1.53	-5.0	3.8	50	0.77	1.60	-3.2	5.2
Gasoline	170	4.17	14.76	-37.7	32.6	110	5.13	12.96	-33.9	24.4	50	0.78	18.80	-37.7	32.6

Source: Emery and Guo (2019) and CANSIM database (Consumer Price Index, annual average, not seasonally adjusted. Table: 18-10-0005-01, formerly CANSIM 326-0021).

4. Result

PLS model by SIMCA 14.1 software is used for regression of following equation.

$$\Delta Bias = \alpha + \sum_{i}^{N} \beta_i \Delta CPI_i + \varepsilon$$

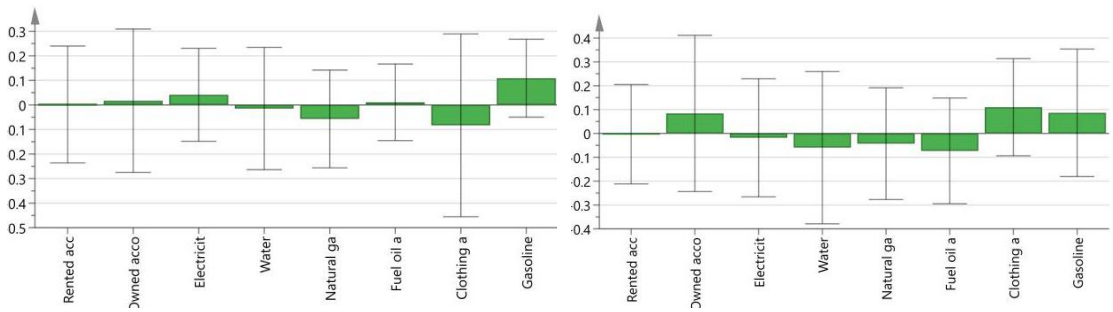
Where, $\Delta Bias$ is the difference term of CPI bias, and ΔCPI is the difference term of the independent variable commodity CPI; α, β is the parameter to be estimated, and ε is the disturbance term. We are concerned about the degreed of influence of each variable, so only the coefficient graph and VIP ranking table of each regression group are listed here based on the PLS method. The significance degree of influence is mainly explained by VIP, while the coefficient diagram mainly shows the direction of influence. If the VIP value exceeds 1, it means that it is very important; if the value is 1-0.5, it means that it is important or the gray area; if the value is below 0.5, it means that it is not important. Therefore, this paper only lists the ingredients that exceed 1.

Urban and rural

It can be seen that the urban group is only affected by electricity in the early stage, while in the later stage it is mainly affected by gasoline,clothing and fuel. The rural group is greatly affected by clothes and natural gas, and the early stage is also affected by gasoline. It is worth noting that for urban, the coefficient of clothing is negative, which indicates that the decline of clothes CPI will lead to the increase of bias. This may be because clothes are low-end consumer goods, and falling prices of clothes will allow

more income to be spent on other consumer goods, which will bring more bias. The negative coefficient of natural gas and fuel may be due to the substitution effect between domestic fuels.

Figure 4.1 Regression coefficient of CPI bias in the groups of urban and rural



Urban Rural

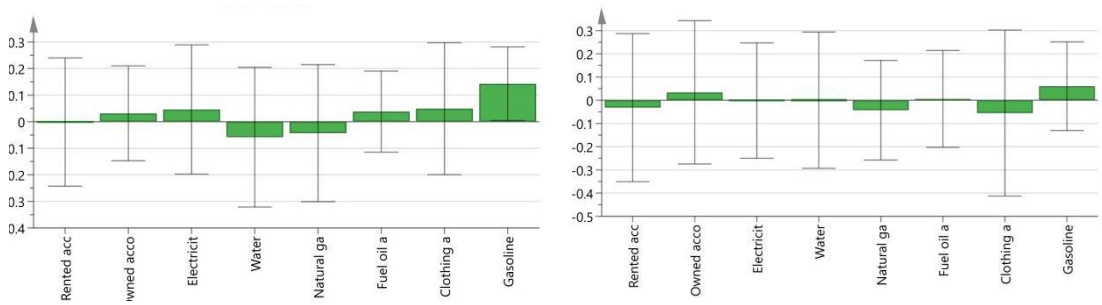
Table 4.1: VIP ranking of independent variables in the groups of urban and rural

Period	Group	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline
1999-2015	Urban						3	2	1
	Rural				2			1	
1999-2009	Urban			1					
	Rural					1		3	2
2011-2005	Urban						3	2	1
	Rural					1		2	

Male and female

It can be seen that male are affected by clothing and owned accommodation in the early stage, while in the later stage, they are affected by gasoline and fuel. In the early stage, female are affected by water, electricity and clothes, while in the later stage, they turn to gasoline, fuel, natural gas and other fuels.

Figure 4.2 Regression coefficient of CPI bias in the groups of male and female



Male Female

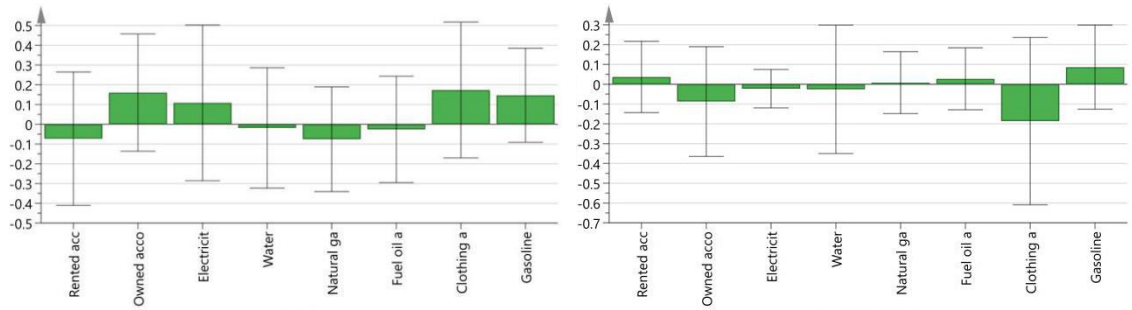
Table 4.2: VIP ranking of independent variables in the groups of male and female

Period	Group	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline
1999-2015	Male						2		1
	Female						3	2	1
1999-2009	Male		2					1	
	Female			1	2			3	
2011-2005	Male						2		1
	Female					1	3	4	2

Married and not married

It can be seen that there is a small difference between the married group and the non-married group at the early stage, both of which are greatly affected by electricity and gasoline. The latter two groups turn to gasoline, fuel, natural gas and other fuels. But the unmarried group is significantly affected by clothing.

Figure 4.3 Regression coefficient of CPI bias in the groups of married and not married



Married Not married

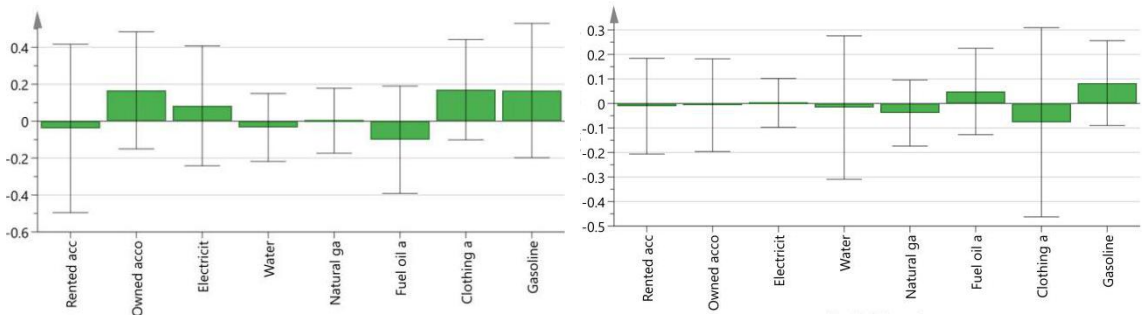
Table 4.3: VIP ranking of independent variables in the groups of married and not married

Period	Group	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline
1999-2015	Married		2					3	1
	Not married						3	1	2
1999-2009	Married			1					2
	Not married	4		1		2			3
2011-2005	Married					2	3		1
	Not married						3	1	2

With children and without children

In the early stage, these two groups are most affected by electricity. The group with children is also affected by natural gas and fuel, while the group without children is affected by owned accommodation and clothes. In the later stage, both move to gasoline, fuel and clothing, but the group without children is still affected by owned accommodation.

Figure 4.4 Regression coefficient of CPI bias in the groups of with children and without children



With children Without children

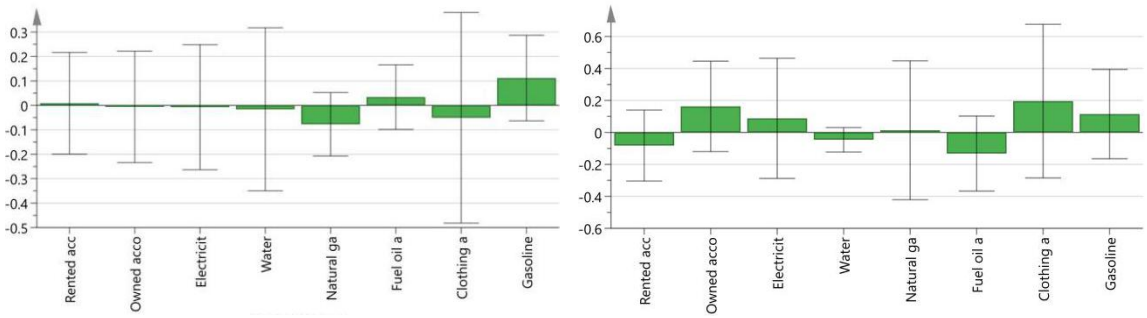
Table 4.4: VIP ranking of independent variables in the groups of with children and without children

Period	Group	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline
1999-2015	With children		1				4	3	2
	Without children						3	2	1
1999-2009	With children			1		2	3		
	Without children		2	1				3	
2011-2005	With children						2	3	1
	Without children		4				3	1	2

Low income and high income

The difference between low-income people and high-income people is larger in the early stage. High-income people are only affected by electricity, while low-income people are affected by clothes, water, electricity and rent. In the later stage, both are affected by gasoline, fuel and clothing.

Figure 4.5 Regression coefficient of CPI bias in the groups of low income and high income



Low income High income

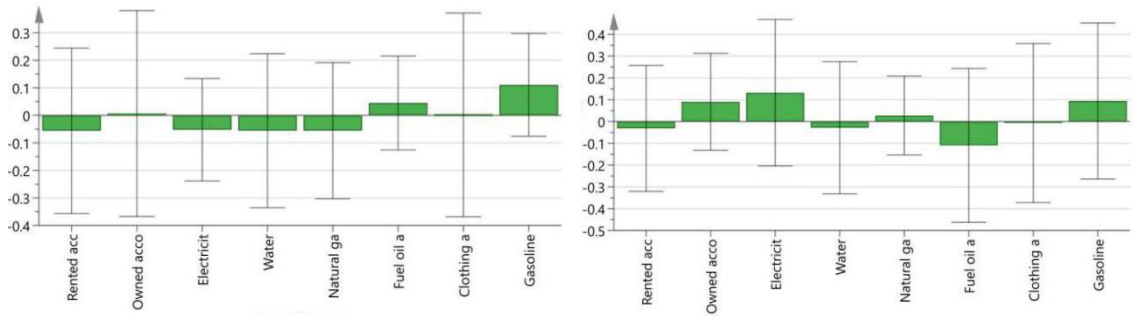
Table 4.5: VIP ranking of independent variables in the groups of low income and high income

Period	Group	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline
1999-2015	Low income					3		2	1
	High income		2				3	1	
1999-2009	Low income	4		3	2			1	
	High income			1					
2011-2005	Low income						3	2	1
	High income						3	2	1

Seniors and not seniors

In the early stage, the seniors are only affected by electricity, while the non-seniors are also affected by rented accommodation. In the later stage, both are affected by gasoline, while the seniors are also affected by natural gas, fuel and owned accommodation, while the non-seniors are also affected by clothing.

Figure 4.6 Regression coefficient of CPI bias in the groups of seniors and not seniors



Seniors Not seniors

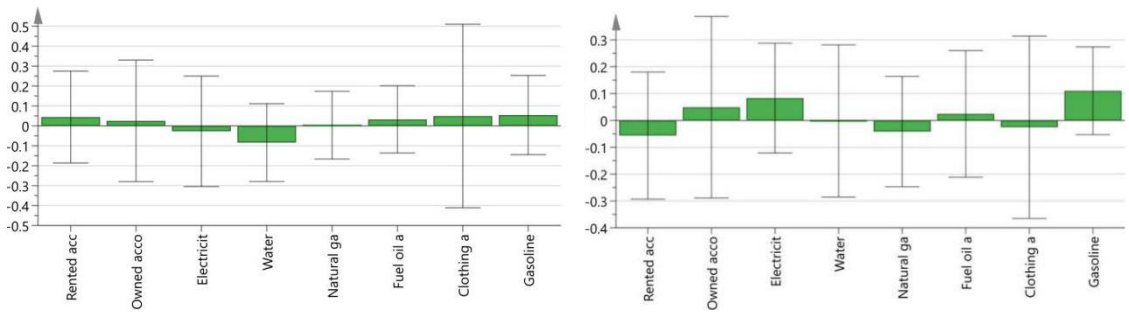
Table 4.6: VIP ranking of independent variables in the groups of seniors and not seniors

Period	Group	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline
1999-2015	Seniors						2		1
	Not seniors		2	1			3		
1999-2009	Seniors			1					
	Not seniors	2		1					
2011-2005	Seniors		4			2	3		1
	Not seniors							2	1

Smokers and non-smokers

Smokers are affected by natural gas and rented accommodation, while non-smokers are only affected by electricity. Later on, both are affected by gasoline and fuel, while non-smokers are also affected by clothing.

Figure 4.7 Regression coefficient of CPI bias in the groups of smokers and non-smokers



Smokers Non-smokers

Table 4.7: VIP ranking of independent variables in the groups of smokers and non-smokers

Period	Group	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline
1999-2015	Smokers				1	4	3		2
	Non-smokers			2					1
1999-2009	Smokers	2				1			
	Non-smokers			1					
2011-2005	Smokers						1		2
	Non-smokers						3	2	1

CPI bias of most groups is affected by gasoline, fuel oil and other fuels and clothes. It can be seen that in the early stage, most groups are affected by electricity, while in the later stage, most groups are affected by gasoline and fuel. However, in the total period, the electricity is not significant, and gasoline and fuel are significant, indicating that the later stage has greater influence and drives the results of the whole sample. On the whole, the differences of different components are relatively small, some differences can be found in the early stage, and almost all of them shift to gasoline, fuel and clothes in the later stage, and there is almost no difference. The difference between early and late stage of each component is large.

Considering the exchange rate

Considering that prices are affected not only by domestic commodity prices, but also by imported commodities. Therefore, we add the first-order difference of Canadian dollar exchange rate into the original regression as an independent variable. The data come from the international monetary fund (IMF). Table 4.8 shows the ranking results of VIP after adding exchange rate. It can be seen that the exchange rate has a significant impact on CPI bias, it could reflect that the official CPI weights do not pick up the impact of the exchange rate through import prices. And the effect was stronger in the later stages than in the early stages.

Table 4.8: VIP ranking of independent variables with considering the exchange rate

Period	Group	Rented	Owned	Electricity	Water	Natural gas	Fuel	Clothing	Gasoline	Exchange rate
	Urban							3	2	1
	Rural				3		4	1		2
	Male						3		1	2
	Female						3	4	2	1
	Married		2				4	3	1	
	Not married						4	1	3	2
1999-2015	With children		1	4			5	3	2	
	Without children						4	3	1	2
	Low income						3		2	1
	Highincome		2					1	3	
	Seniors						2		1	3
	Not Seniors		3	1					2	4
	Smokers				3	5	4		2	1
	Non-smokers			3					1	2
	Urban			1						
	Rural					1		4	3	2
	Male		2					1		
	Female			1	2			3		
	Married			1					3	2
	Not married			2		1				3
1999-2009	With children			1		2	4			3
	Without children		2	1				3		
	Low income	4		3	2			1		
	Highincome			1						
	Seniors	3				1	2		5	4
	Not Seniors	2		1						
	Smokers	3				1				2
	Non-smokers			1						
	Urban						4	3	1	2
	Rural					2	3			1
	Male						3		2	1
	Female					2	4		3	1
	Married						3		2	1
	Not married						3	1	2	
2011-2015	With children						3		2	1
	Without children		5				4	1	3	2
	Low income						4	2	3	1
	Highincome						2	4	3	1
	Seniors					4	3		2	1
	Not Seniors						2	4	3	1
	Smokers						1		2	3
	Non-smokers						3	4	2	1

5. Conclusion

Since the official CPI does not take into account the consumption preferences and consumption weights of different groups, there will be bias in measuring the cost of living of different groups. I have used the partial least squares method to explore the influence of CPI changes of different commodities on CPI bias of 7 sub-groups in the Canadian population. The results showed that the causes of CPI bias at the early stage of each group are significantly different. However, after 2010, in the post-financial crisis era, the causes of CPI bias of each group have a similar trend, namely, focusing on gasoline, clothes and fuel. After considering the exchange rate, the exchange rate also becomes an important factor affecting the CPI bias, suggesting that CPI bias may also be affected by imported goods. This effect is more important at a later stage. The differences in the impact of CPI in different groups reflect the consumption structure characteristics and cost of living components of different groups, which should be taken into account when formulating differentiated tax policies and welfare policies.

PLS method is a method used to deal with independent variables multicollinearity, which is suitable for the characteristics of independent variables in this paper. However, for the purpose of this paper, we are mainly concerned with the impact of CPI changes of each commodity, rather than constructing a theoretical model to fully explain CPI bias, so some PLS-related indicators are not applicable. Further work is needed to find more explanatory variables to filter and construct a model of CPI bias. In this way, we can

verify the accuracy of the whole model by using PLS relevant indicators, and predict the CPI bias.

6. Reference

Bruce W. Hamilton. (2001a). Using Engel's Law to Estimate CPI Bias. *American Economic Review*, 619-630.

Bruce W. Hamilton. (2001b). Black-White Differences in Inflation: 1974-1991. *Journal of Urban Economics* 50, 77-96.

Chen, Q. (2014). Advanced econometrics and Stata applications. Adv. Educ. *Publishing House*.

Dora Costa. (2001). Estimating Real Income in the United States from 1888 to 1994: Correcting CPI Bias Using Engel Curves. *Journal of Political Economy*, 109 (6), 1288-1310.

Greene, W. H. (2003). Econometric analysis. *Pearson Education India*.

Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research: a review of four recent studies. *Strategic management journal*, 20(2), 195-204.

Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage*, 56(2), 455-475.

Matthew Brzozowski. (2006). Does One Size Fit All? The CPI and Canadian Seniors. *Canadian Public Policy / Analyse de Politiques*, 32 (4), 387-411.

Monk, W. A., Wilbur, N. M., Curry, R. A., Gagnon, R., & Faux, R. N. (2013). Linking landscape variables to cold water refugia in rivers. *Journal of environmental managemen*t, 118, 170-176.

Pistonesi, M. F., Di Nezio, M. S., Centurión, M. E., Palomeque, M. E., Lista, A. G., & Band, B. S. F. (2006). Determination of phenol, resorcinol and hydroquinone in air samples by synchronous fluorescence using partial least-squares (PLS). *Talanta*, 69(5), 1265-1268.

Wold, S., Albano, C., Dunn, M., Esbensen, K., Hellberg, S., Johansson, E., & Sjöström, M. (1983). Pattern regression finding and using regularities in multivariate data. *Analysis Applied Science Publication*, London.

Curriculum Vitae

Candidate's full name: Haoyu Wang

Universities attended: Bachelor of Finance

Jilin University, 2015.6.

Master of Economics

Shandong University of Finance and Economic, 2018.6.

Publications: N.A.

Conference Presentations: N.A.