CE-Based Consumer Expenditure Behaviors Study

Zhicong Zhao(George)*

Fuyuan Li(David)†

Abstract
In this study, we use varying coefficient model to describe the relationship between household age and hump-shaped consumer expenditure. To demonstrate the benefit of varying coefficient model, a comparison has been done with two traditional models via regression results and parameters visualization. This methodology is applied on, but not limited to, beef expenditure in pubile-use microdata.

Key Words: Varying Coefficient Model, Hump-shaped consumer expenditure, Regression

1. Introduction

In recent study, consumer expenditure (CE) public-use microdata (PUMD) is being widely used in economics. Based on the high quality and long-time period it covers, we can easily obtain household information and expenditure records by directly using PUMD. For example, Attanasio and Weber (1995) used CE data for consumption growth study and Attanasio and others (1999) showed hump-shaped lifetime consumption study via CE data. However, in previous studies (Attanasio and Weber, 1995; Attanasio et al., 1999), people would rather define cohort by 5-year birth interval instead of every single year. One of the main reasons is that splitting PUMD by every single year birth leads to huge variances in average expenditures, and those extreme value will significantly affect regression results, like the comparison shows in figure 1. Plot on left side represents the average expenditure by 5-year birth interval and plot on right side represents the average expenditure by 1-year birth interval. Even with similar main trend, 5 years period cohort gives us a stable pattern, while single year period cohort gives us more details. An accurate single year cohort model is valuable if we can figure out an effective statistics method to eliminate the effects for extrema. In our study, varying coefficient model as a functional extension of ordinary linear models (Fan and Zhang, 1999) is applied on PUMD dairy beef expenditure to establish single year cohort expenditure model. Based on the limitation of our knowledge, we are the first one who try to use varying coefficient model on PUMD for single year period cohort expenditure study. All the expenditure information comes from file “expd” and household profile information comes from file “fmld”.

2. Three Models

During data preparation, the data (from 1996 to 2015) is grouped by household brith year. Considering validation of sample size for each cohort, household who were born from 1931 to 1980 are the concern in this study.

Since expenditure profile is hump-shaped, a quadratic trend in regression model can be an appropriate basic assumption. Thus, set

$$Y_m = a(m)X^2 + b(m)X + c(m) + \varepsilon_m \tag{1}$$

$m = 1, 2, ..., 50$ is cohort index. $m = 1$ means we are dealing with household who were born in 1931, $m = 2$ represents cohort who were born in 1932 and so on so forth.

X is a column vector represent household’s age. Since the data is uploaded every quarter year, we add .00, .25, .50, .75 when generating X in each season respectively. For example, household who were born in 1980 is marked as age 16.00 in 1996 first season, marked as age 16.25 in 1996 second season, 16.50 in 1996 third season and 16.75 in 1996 fourth season. Dimension of X is 80×1 for 20 years modeling period.

[*] Mississippi State University, B. S. Hood Rd, Mississippi State University, MS 39759
[†] The George Washington University,2121 I St NW, Washington, DC 20052

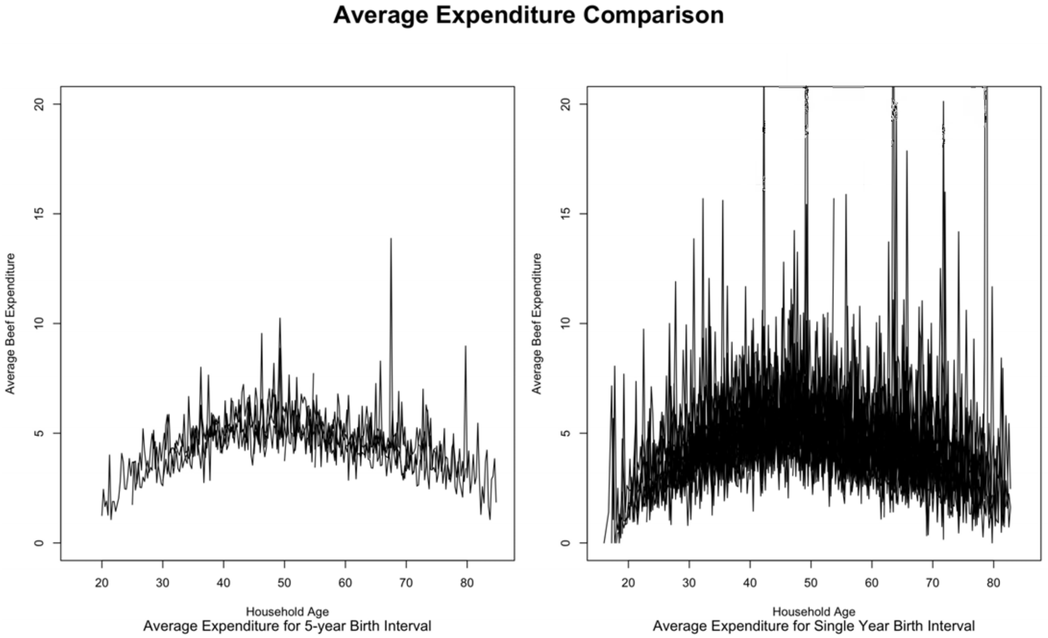

Figure 1: 5-year Birth Interval & 1-year Birth Interval.

Y_m is a column vector which represents cohort m's weekly average expenditure with respect to age. Dimension of Y_m is also 80×1.

ε_m is a random vector with dimension 80×1 and each element has independent $N(0, \sigma_m^2)$ distribution.

2.1 Simple Regression Model

Starting with the simplest case, we assume all the parameters in different cohorts are constants. In this case, the model assumptions reduce to below:

$$Y_m = aX^2 + bX + c + \varepsilon_m,$$

ε is a random vector with all elements are independent and identical $N(0, \sigma^2)$ distributed and a,b, c and σ^2 are unknown constants.

The table 1 shows the results of simple regression parameters.

Table 1: Parameter Estimations from Simple Regression Model with Full Data from PUMD

â	$\hat{b}$	ĉ
-0.003	0.293	-1.872

As figure 2 shows, red quadratic line represents predictions and grey segments are observed average expenditures. Since all the cohort share a same set of parameters, the prediction line is compact and simple. Based on the model assumption, this line can reflect the hump shape in expenditure but limited in indicating expenditure pattern changes by generation.

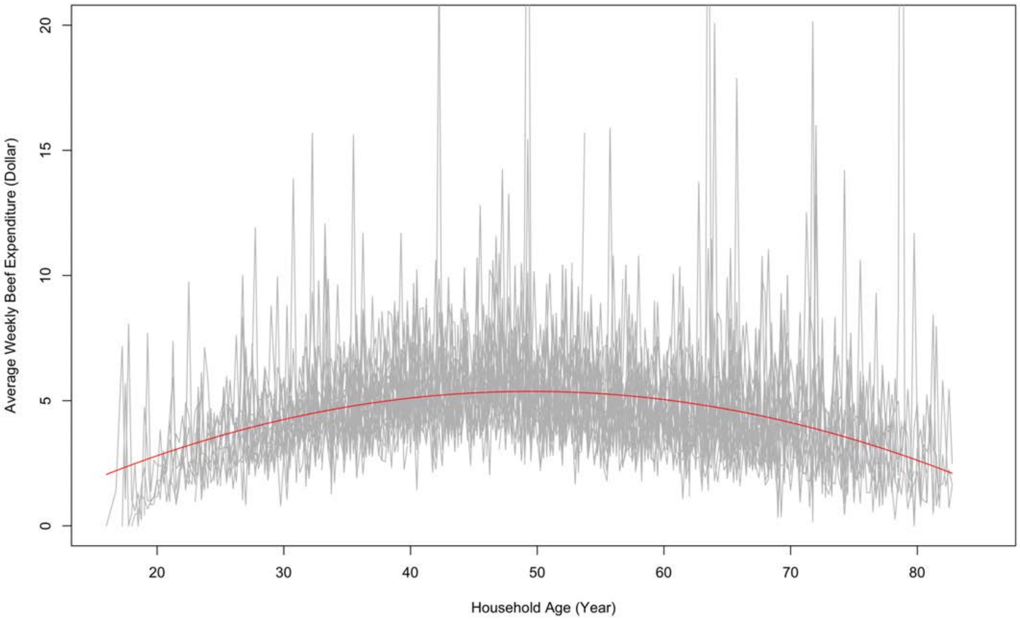

Figure 2: Predictions for Simple Regression Model

2.2 Cohort Based Regression Model

In cohort based regression model, the data is split by every single cohort. Thus, the correlation among different cohorts are completely ignored. Each regression model is built up with respect to data only in corresponding cohort. In this way, we can relax the model restrictions as follows:

$$Y_m = a(m)X^2 + b(m)X + c(m) + \varepsilon_m$$

ε_m is a random vector with all elements independent $N(0, \sigma_m^2)$ distributed and $a(m)$, $b(m)$, $c(m)$ and σ_m^2 are functions with respect to m.

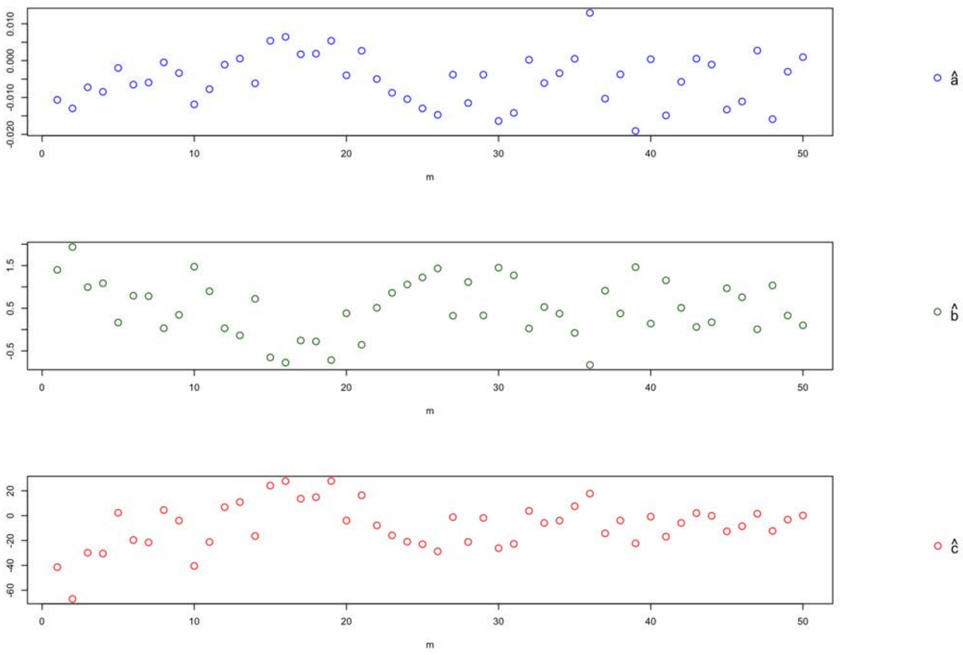

Figure 3: Parameters for Cohort Based Regression Model.

The figure 3 contains three subplots. Each subplot shows the empirical parameter function for $a(m)$, $b(m)$, $c(m)$, respectively. By individual subplot, we can see the way individual parameter changes by cohort or generation. Additionally, points with same cohort index in all three subplots reflect the regression coefficients for each cohort. Comparing with simple regression model, this model gives more flexibility in parameters and allow us to observe the trend hiding among different generations for expenditure pattern. However, the affection of extreme values takes away from model accurateness. In figure 4, grey lines are observations from the full PUMD and red curves represent expected regression lines for fifty cohorts. Some of the prediction lines are deviate from the main trend due to extrema.

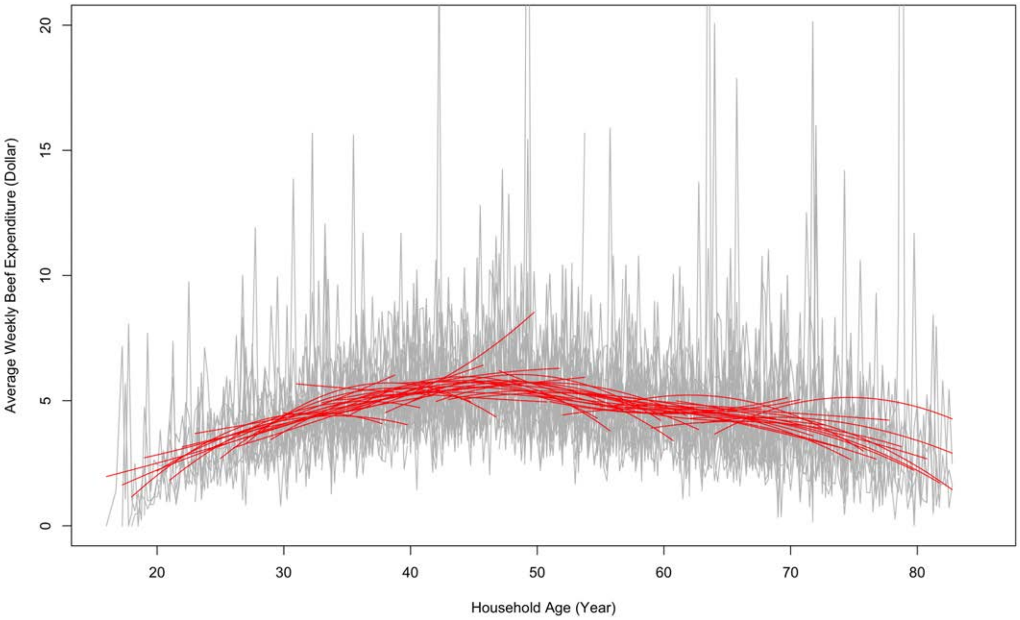

Figure 4: Perdictions for Cohort Based Regression Model.

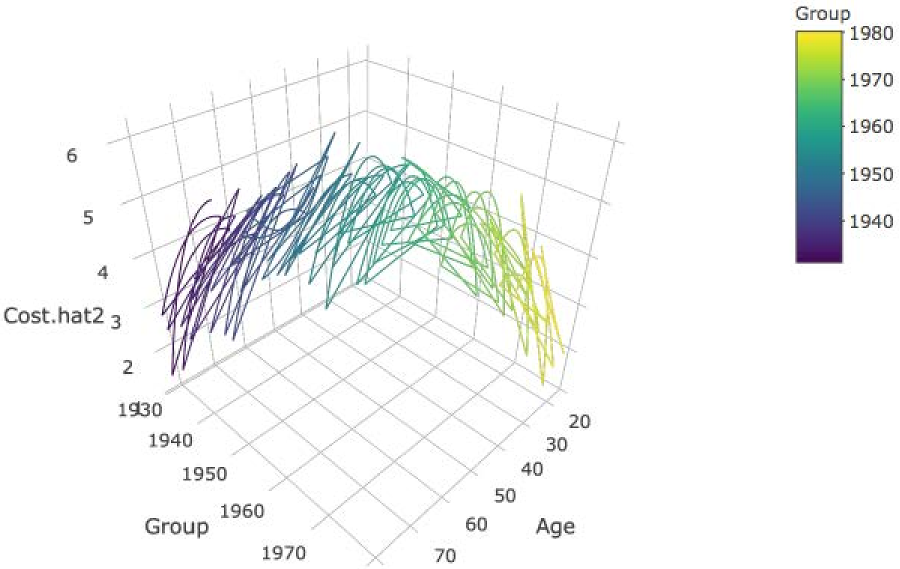

Figure 5: Perdictions for Cohort Based Regression Model 3D Plot.

Figure 5 is a 3D version for figure 4.

2.3 Varying Coefficient Model

The model in section (2.1) is a regression model based on common regressors for all cohorts, while in section (2.2) the regression model has different regression coefficients for each cohort. Intuitively, The Cohort Based (CB) Model assumes people at the same age may have different preference on beef expenditure based on different born year they were. Say, even both are at their 20s, CB model allows people born in 1960s have distinct mean level of beef expenditure from people born in 1990s. Such trend was also detected in Attanasio, Banks, Meghir and weber (1999). The economic factors that drive such different preferences are out of the scope of this paper, but a model that accounts such trends is more realistic and is more scientific meaningful.

Even though the CB Model outperforms the regression model with common regressors, which is expected since model (2.2) was based on each cohort and has more accurate estimators, there are still some limitations for Cohort Based Model.

First, by treating each cohort separately, such model lose information in the sense that it actually involves a small portion of the whole dataset. Development of modern techniques makes the cost of data collection low, even of massive amounts. Nevertheless, in CB model, we cannot take advantage of big data. Even with more data collected, more cohorts will be involved, the increasing of numbers of observations is limited.

Second, the interpretability is problematic for CB model. We treat each cohort separately in CB model and hence the estimators for each cohort are not contributing to the interpretation of other cohorts. Intuitively, however, one may expect impacts from ”neighborhoods” for a certain cohort's preference on beef expenditure. The CB model is somehow weak in preserving such property.

To account for the above two issues, we propose to apply the varying coefficient model on CE data. Varying coefficient model was first introduced by Hastie and Tibshirani (1993), where they used such model to analyze a survival data set. Fan and Zhang (1999) proposed a two-step estimator for varying coefficient model and studied their asymptotic properties. They also showed the improvements of two-step varying coefficient model estimator by applying to an environmental data set. As our best knowledge, this paper is the first to apply such varying coefficient model to CE data set to study consumers' expenditure behaviors.

In varying coefficient model, we treat our model coefficients as some smoothed functions of cohorts. By using kernel based smoothing estimation, we actually involves all data points even for estimation on one specific cohort, which avoids the information loss since the smoothing process will take all the ”neighbor” cohort into account.

Another motivation for choosing varying coefficient model is that it preserves the dynamic trends in data set (Fan and Zhang, 2008). For instance, the mean level and (linear) age impact to beef expenditure among each cohort, although may not be constant, will not be expected to change drastically. Instead, by assuming some smoothed function on cohort, information can be borrowed from neighbor cohorts. In our analysis on CE data, the cohort is based on born year. By assuming varying coefficient model, therefore, we are assuming that the mean level of preference on beef expenditure for 1950s at their 30s, will be affected more or less by the preference of people of 1960s and 1940s at their 30s. Even though our model will not dive into the economical or psychological factors behind such behavior and impact, revealing such trend from statistical perspective is the scientific contribution from our analysis.

Recall the equation (1), In varying coefficient model, we will further assume the estimator $a(m)$, $b(m)$ and $c(m)$ are some smoothed function of m, $A(\cdot)$, $B(\cdot)$ and $C(\cdot)$.

$$Y_m = A(m)X^2 + B(m)X + C(m) + \varepsilon_m,$$

where we observe (M_i, X_i, y_i), where $M_i = 1931 \ldots 1980$ is the cohort label, X_i is the age of each observation and y_i is the beef expenditure, and $E(\varepsilon) = 0$, $var(\varepsilon) = \sigma^2(M)$.

Hence, for each given m, the estimator $\{\tilde{A}(m), \tilde{B}(m), \tilde{C}(m)\}$ of $\{A(m), B(m), C(m)\}$ can be obtained by solving

$$L(a,b,c) = \Sigma_{i=1}^{n}(y_i - aX_i^2 - bX_i - c)^2 K_h(M_i - m), \tag{2}$$

where $K_h(s) = K(s/h)/h$ and $K(s)$ is a kernel function $K(s) = \phi(s)$. In our analysis we choose Gaussian kernel, while some other choice like Epanechnikov kernel $K(s) = 0.75(1-s)_+^2$ can be considered (Fan and Zhang, 2008).

Note here that in the kernel based smoothing function, we need to decide the bandwidth h. To obtain the optimal bandwidth, we apply the cohort-based cross validation.

1. choose bandwidth h_1 among all candidate bandwidth $h_1, \ldots h_k$.

2. for m ranges from 1931 to 1980, do the followings:

(a) remove mth cohort and use subset of data (M_i, X_i, y_i) with $M_i \neq m$ to build model, by minimizing the loss function (2).

(b) use the estimated smoothed parameters to predict $\hat{y}_i$ based on (M_i, X_i) with $M_i = m$.

$$\hat{y}_i = \tilde{A}(m)X_i^2 + \tilde{B}(m)X_i + \tilde{C}(m)$$

(c) calculate the MSE of prediction on cohort m, mse_m.

3. add up all the m MSE, mse_{1931} to mse_{1980} to obtain the cross-validation error $CV.e_1 = \Sigma_{m=1931}^{1980} mse_m$, which is corresponding to bandwidth h_1.
4. repeat step 1 to step 3 for other candidate bandwidth and obtain corresponding $CV.e_2$ to $CV.e_k$.
5. the optimal bandwidth is the h_{opt} gives the smallest $CV.e_{opt}$.

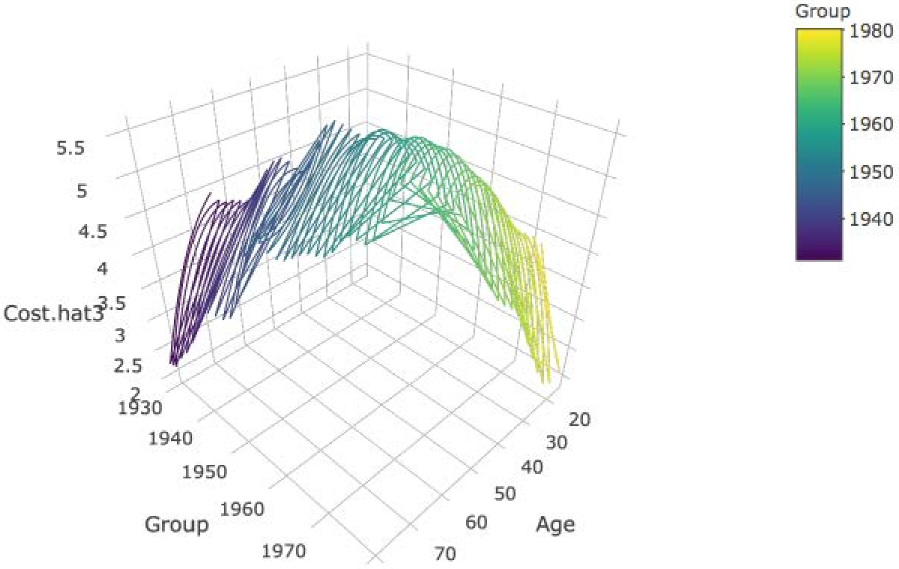

Figure 6: Perdictions for Varying Coefficient Regression Model 3D Plot.

Figure 6 is 3D plot for prediction, compare with figure 5, this one is well ordered and concise.

3. Models Comparison

The parameters visualization demonstrates the advantages of varying coefficient model against traditional regression methods. In this section, we will use cross validation to compare the robustness and accuracy for all three models. Cohort $m = 20$ has relatively less extreme values than others so that this cohort could be treated as the valid sample. Thus, cross validation only has been done on $m = 20$.

First data from cohort $m = 20$ has been removed, which is household who were born in year 1950. Second, we use three models fit with the data left in other 49 cohorts respectively. There are three sets of predictions of $a(20)$,$b(20)$ and $c(20)$ from three different models. The comparison is done via model predictions and observations in cohort $m = 20$. The subscript s represents results from simple regression model, subscript c represents results from cohort based regression model and subscript vc represents results from varying coefficient model.

For the simple regression model, after removing cohort with index $m = 20$, the results of parameters are shown in table 2 The parameters are exactly same with the simple regression model applied

Table 2: Parameter Estimations from Simple Regression Model without Cohort Were Born in Year 1950 PUMD

â	b̂	ĉ
-0.003	0.293	-1.872

Table 3: Original Parameter Estimations from Cohort Based Regression Model without Cohort Were Born in Year 1950 PUMD

m	...	19	20	21	...
$\hat{a_c}$	...	0.005	NA	0.003	...
$\hat{b_c}$	...	-0.714	NA	-0.355	...
$\hat{c_c}$	...	27.865	NA	16.368	...

on full dataset when round up to three decimal places. The results indicates that simple regression model is robust.

For the second model, since the correlation among different cohorts are ignored, the $a(20)$,$b(20)$ and $c(20)$ can not be directly estimated by data other cohorts. However, if $a(20)$,$b(20)$ and $c(20)$ are treated as missing values in their own parameter functions, the average of parameters in previous cohort and next cohort are reasonable approximation. Table 3 above is the original estimations from data. Then we use average of adjacent values to replace the missing values and table4 shows as below.

As for varying coefficient model, $a_{vc}\hat{(}20)$, $b_{vc}\hat{(}20)$ and $c_{vc}\hat{(}20)$ comes from

$$\arg\min_{a,b,c} \sum_{m=1,m\neq 20}^{50} \sum_{i=1}^{n_m} (y_{mi} - ax_{mi}^2 - bx_{mi} - c)^2 * K(\frac{20-m}{h})$$

n_m is number of different ages in cohort m.

y_{mi} and x_{mi} are the weekly average expenditure for cohort m at age i.

$K(\frac{20-m}{h})$ is a kernel density function and here we use Gaussian kernel density function.

After optimized, the parameters prediction shows in table 5.

Figure 7 is drawn by previous regression parameters. In the picture, red line represents predictions come from simple linear regression, green line represents predictions come from cohort based linear regression and blue line represents predictions come from varying coefficient model. By visualization, the performance of three lines are similar. However table 6 shows that their SSR are slightly different and varying coefficient model is the one with the minimum SSR.

4. Conclusion

The constant simple regression model and cohort based regression model both have its own advantages. The former one can give us stable estimations but unable to indicate how parameters change as time pass by. On the other hand, the latter one allows parameters change by generation but the extreme values highly affect the veracity of parameter estimations. In this case, those two models are limited and unrecommended.

Table 4: Parameter Estimations from Cohort Based Regression Model without Cohort Were Born in Year 1950 PUMD With Average Fit in Missing Part

m	...	19	20	21	...
$\hat{a_c}$	...	0.005	0.004	0.003	...
$\hat{b_c}$	...	-0.714	-0.535	-0.355	...
$\hat{c_c}$	...	27.865	22.117	16.368	...

Table 5: Parameter Estimations from Varying Coefficient Model without Cohort Were Born in Year 1950 PUMD

$\hat{a_{vc}(20)}$	$\hat{b_{vc}(20)}$	$\hat{c_{vc}(20)}$
-0.001	-0.166	11.632

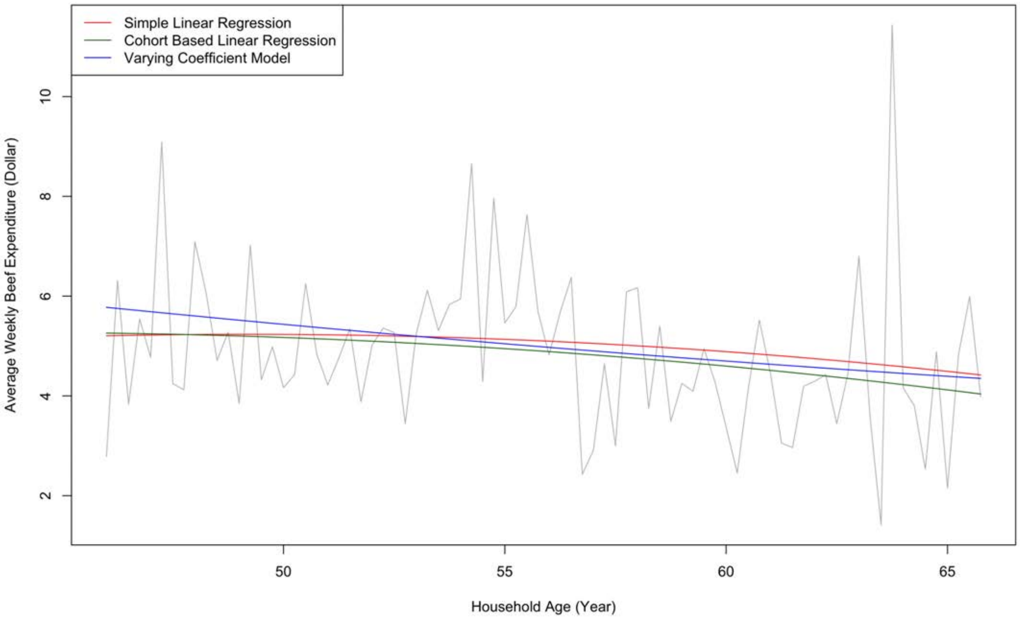

Figure 7: Models Comparison after Removing Cohort 20.

Meanwhile, that is the reason why varying coefficient model is an appropriate model. Even it is similar with cohort based regression model, but in fact, global data is used instead of individual cohort data when fitting with the model. The kernel density function of each cohort could guarantee the household who were born closer to response cohort can make more contribution in regression model than household who were born far away from this cohort. Thus, varying coefficient model proved smoother and more accurate and robust regression parameters than the first two models.

Table 6: SSR Comparison for Simple Regression Model, Cohort Based Regression Model and Varying Coefficient Model

SSR for Simple Regression Model	SSR for Cohort Based Regression Model	Varying Coefficient Model
201.1758	199.4055	196.3129

5. References

References

[1] Attanasio, O. , Banks, J. , Meghir, C. & Weber, G. (1999) Humps and Bumps in Lifetime Consumption, Journal of Business & Economic Statistics, 17:1, 22-35

[2] Attanasio, O. & Weber, G.(1995). Is Consumption Growth Consistent with Intertemporal Optimization? Evidence from the Consumer Expenditure Survey. Journal of Political Economy Volume 103, Number 6 (Dec.,1995)

[3] Fan, J. & Zhang, W.(1999). Statistical Estimation in Varying Coefficient Models.The Annals of Statistics Volume 27, Number 5 (1999), 1491-1518.

[4] Fan, J. & Zhang, W.(2008). Statistical Methods with Varying Coefficient Models.Statistics and Its Interface Volume 1, (2008), 179?195.

[5] Hastie, T. & Tibshrani, R.(1993). Journal of the Royal Statistical Society. Series B (Methodological) Vol. 55, No. 4 (1993), pp. 757-796.

[6] Fernández-Villaverde, J.& Krueger, D.(2007). Consumption over the Life Cycle: Facts from Consumer Expenditure Survey Data. The Review of Economics and Statistics. 89(3):552-565. doi: 10.1162/rest.89.3.552.

[7] Pan A, Sun Q, Bernstein AM, Schulze MB, Manson JE, Stampfer MJ, Willett WC, Hu FB. Red Meat Consumption and MortalityResults From 2 Prospective Cohort Studies. Arch Intern Med. 2012;172(7):555-563. doi:10.1001/archinternmed.2011.2287