SHIFT FUNCTION PLOTS FOR REGRESSION FITTING

Zilin Wang[1] and David R. Bellhouse[2]

ABSTRACT

Although a nonparametric regression model allows us to obtain a graphical display of the relationship between the response variable and the independent variables, the exact form of the regression function is not evident. This weakness can be overcome by a parametric model since the relationship between the dependent and the independent variables is specified mathematically. The objective of this paper is to introduce a new graphical approach, called the shift function plot, with which a hypothesis test is constructed to evaluate the goodness of fit of a parametric regression model. Under multi-stage stratified sampling schemes, we apply the shift function plots to survey data. Asymptotic properties of the survey estimator of the shift function are established. An empirical example from the 1990 Ontario Health Survey is used to illustrate the application of the shift function.

KEY WORDS: Complex surveys, Nonparametric regression; Probability sampling; Regression analysis; Regression specification test.

RÉSUMÉ

Bien que les modèles de régression non paramétrique nous permettent d'obtenir des représentations graphiques de la relation entre la variable réponse et les variables indépendantes, la forme exacte de la fonction de régression n'est pas toujours évidente. Cette faiblesse peut être surmontée par un modèle paramétrique puisque la relation entre les variables dépendantes et indépendantes est spécifiée mathématiquement. Ici nous présentons une nouvelle approche graphique appelée le graphique de la fonction de décalage à partir duquel un test d'hypothèse est construit pour évaluer la qualité de l'ajustement d'un modèle de régression paramétrique. Sous un plan d'échantillonnage stratifié à plusieurs degrés, nous appliquons les graphiques de fonction de décalage à des données d'enquête. Les propriétés asymptotiques de l'estimateur de fonction de décalage de l'enquête sont ensuite établies. Un exemple empirique tiré de l'Enquête sur la santé en Ontario de 1990 est utilisé afin d'illustrer l'application de la fonction de décalage.

MOTS CLÉS : Analyse de régression; échantillonnage probabiliste; enquêtes à plan complexe; régression non paramétrique; test de spécification de régression.

1. INTRODUCTION

To avoid the problem of misspecification in regression analysis, nonparametric regression techniques, also called smoothing, have become very commonly utilized to conduct flexible modelling. When estimating a regression function nonparametrically, we do not project the observed data into a regression function. Instead, we obtain a point estimate of the regression function at a fixed point of the independent variable. A great deal of research has shown that the use of a smoother results in a more accurate or closer approximation of the regression function to the true model than parametric regression techniques. Although nonparametric techniques are very effective tools for flexible modelling, they do not provide the exact form of the regression function. Even though a parametric regression model risks model misspecification and is more rigid in terms of the modelling, it provides a clear scientific relationship between the independent and dependent variables. Because of this clearer interpretive characteristic, a parametric model is favoured for the occasions on which it is necessary for researchers to have the same underlying "law" reflected in a regression model. In terms of the interpretable purpose and the scientific experiment, parametric regression models do a more promising job than a nonparametric model does. Hence, if there is a way to guarantee accurate modelling, a parametric regression is a more useful way to resolve scientific questions than a nonparametric regression model. The objective of this paper is to develop a graphical diagnostic method for the specification of a parametric regression model with the aid

1 Zilin Wang (zwang@wlu.ca), Department of Mathematics, Wilfrid Laurier University, Waterloo, Ontario, Canada, N2L 3C5.
2 David R. Bellhouse (bellhouse@stats.uwo.ca.), Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada, N6B 5A7,

of a smoother. According to the overview of nonparametric lack of fit tests by Hart (1997, chapter 6), there are two fundamental approaches, in terms of smoothing, involved in nonparametric tests: smoothing the residuals or comparing the parametric and nonparametric models. The rationale of our test is very similar to the use of comparing parametric and nonparametric models, but the test statistics are different. In the aforementioned two approaches, the residuals from the regression fits are smoothed again, whereas in our method we rely only on the properties of the estimated regression functions and the predicted values.

Our method is motivated by the idea of a shift function introduced by Doksum and Sievers (1976). In their work, a shift function is defined as: $\Delta(x) = G^{-1}(F(x)) - x$, where G and F are two continuous distribution functions of two random variables. By examining the estimates of the shift function, $\hat{\Delta}(x) = \hat{G}^{-1}(\hat{F}(x)) - x$, we can compare the distributions of the random variables of interest. Here, we generalize this idea of the shift function for use in hypothesis testing or diagnostic checking of the regression model fits. We establish two kinds of the tests based on the shift function. One is the goodness of fit or specification test and the other is the homogeneity test. In the goodness of fit test, we parameterize the model under the null hypothesis. Using the nonparametric model to estimate the regression function, we can examine how closely the parametric model explains the data. With the homogeneity test, we can compare the regression fit between two populations or two strata. In Section 2, we introduce the methodology and apply the shift function to the hypothesis testing. Sampling framework and working models will is set up in Section 3. Section 4 is devoted to setting up the estimations and establishment of the moments and asymptotic properties of survey estimator of the shift function. An example from the 1990 Ontario Health Survey is used to illustrate the diagnostic method in Section 5. Concluding remarks are presented in Section 6.

2. METHODOLOGY AND HYPOTHESES TESTS

Consider the regression function $y = m(x) + \varepsilon$ where $\varepsilon$is independent and identically distributed with mean 0 and variance σ^2. We can conduct estimation of $m(x)$ with either parametric or nonparametric modelling. When we establish hypothesis tests, we consider the composite hypothesis that,

$$H_0 \colon m(x) = m_0(x; \boldsymbol{\beta}) \tag{1}$$

with the alternative hypothesis being H_1: $m(x) \neq m_0(x;\boldsymbol{\beta})$. The function $m_0(x; \boldsymbol{\beta})$ is a parametric univariate regression function of x where $\boldsymbol{\beta}$ is a $1{\times}p$ vector that parameterizes the regression function under the null hypothesis. A polynomial regression function is an example of such a regression function. Because $m(x)$ can be any function, and does not necessarily have the properties of a distribution function, we can not directly adapt the shift function to build up a test. That is, given that $m(x)$ is the true regression function, $m_0^{-1}(m(x))$ does not provide any useful information on the model fits if both $m_0(\cdot)$ and $m(\cdot)$ are not one-to-one positive functions. To solve the problem, we transform $m(x)$ so that the transformed function has the properties of a distribution function. We assume that $m(x)$ is non-negative and integrable, and the integral of $m(x)$ is invertible for all the x on its domain. The transformation is defined as, for all x □ D:

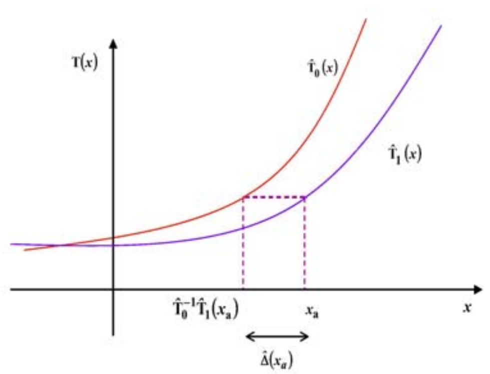

Figure □:A diagram of Shift Function

$$T(x) = \int_a^x m(t)dt$$

Note that domain D is transformed so that □x □ D, $m(x)$ is non-negative and a is the lower bound of D. We define the shift function as:

$$\Delta(x) = T_0^{-1}(T(x)) - x \tag{3}$$

where $T_0^{-1}(\cdot)$ is the inverse of the transformed function under the null hypothesis. Illustration of a shift function can be found in **Figure 1**. The horizontal distance between the projections of $T_0(x)$ and $T_1(x)$ is $\Delta(x)$ and $\Delta(x)$ is zero if $T_0(x)$ and $T_1(x)$ are the same.

Without knowing the true model, it is difficult for us to estimate the shift function defined in (3). With a good nonparametric estimation technique, we can obtain a very reasonable or close approximation of the true model because a smoother is a data driven technique and has been shown to fit data better than a parametric regression model does. Let $\hat{m}_{\mathbf{S}}(x)$ be an estimator of the regression function with a certain type of smoothing technique, where **S** consolidates all the information on the smoothing parameters. **S** also varies according to the smoothing techniques used to estimate $m(x)$. We can estimate the transformation in (1) as $\hat{T}_{\mathbf{S}}(x) = \sum_{i=0}^{[[x-a/\Delta t]]} \hat{m}_{\mathbf{S}}(a + i\Delta t)\Delta t$. Hence, we estimate the shift function with $\hat{\Delta}(x) = \hat{T}_0^{-1}(\hat{T}_{\mathbf{S}}(x)) - x$ where $\hat{T}_0(x) = \int_a^x m(t;\hat{\boldsymbol{\beta}})dt$.

When the null hypothesis is true, the estimated shift function fluctuates around zero. A useful way of using the estimated shift function would be to graph the estimated shift function versus x and see how much, and in what range of x, it differs from zero.

3. SAMPLING FRAMEWORK AND WORKING MODELS

Suppose that we have a population U consisting of N distinct units. The characteristic of interest is a vector-valued unit (y_k, x_k) for all the k =1, …, N. y_k represents the k^{th} population value of the response variable and x_k represents the k^{th} observation of the explanatory variable. Let **s** be the set of units in the sample with (y_k, x_k, w_k) for $k \in \mathbf{s}$ obtained according to the sampling design with sample size n. The survey weight w_k is attached to the k^{th} sampling unit. Additionally, we assume that there is no nonresponse so that the inclusion probabilities are equal to the reciprocals of the sampling weights.

Note that there are several estimation procedures needed to accomplish the estimation of the shift function. In order to obtain estimates and make inferences with them, we assume of a nested sequence of finite populations. The sequence is indexed of by ν such that all the finite population quantities and the sample quantities depend on the index ν and all the asymptotic results are established as $\nu \to \infty$. Additionally, we assume that the N finite population units are a sample of independent and identically distributed units from the infinite superpopulation. The units of a finite population are realization of two working models under the null hypothesis stated in (1). The working models are:

Model 1: The units of a finite population are realization of the model $y=\zeta(x)+\varepsilon$, where ε are independent and identically distributed and $\zeta(x)= E(y|x)$.

Model 2: The units of a finite population are the realization of the model $y = \zeta_0(x;\, \boldsymbol{\beta})+\varepsilon$, where $\zeta_0(x;\boldsymbol{\beta})$ is the parametric regression function and $\boldsymbol{\beta}$ are the vector of the parameter.

We define $m(x)$ and **B** to be the finite population parameters whose superpopulation counterparts are $\zeta(x)$ and $\boldsymbol{\beta}$.

4. ESTIMATION AND ASYMPTOTICS

Using design-based least squares and local polynomial regression techniques, we estimate the model shift function and denote the census estimate by the finite population parameter $\Delta(x)$ and the corresponding survey estimate by $\hat{\Delta}(x)$. Using Model 1 as the working model under the null hypothesis (1), and based on the development of design-based regression analysis in Fuller (1975) and references therein, the survey estimate of the finite population parameter, **B**, is in the form of

$$\hat{\mathbf{B}} = (\mathbf{X}(x)^T \mathbf{W}\mathbf{X}(x))^{-1}\mathbf{X}(x)^T \mathbf{W}\mathbf{y} \tag{4}$$

where $\mathbf{X}(x)$ is an $N\times p$ matrix composed of vectors of parametric functions, such as polynomials of degree p, of x, **y** is the vector of response variable, and **W** is a weight matrix whose diagonal entries are survey weights attached to observations.

One of the characteristics of complex survey data is that the scale of the data set can be very large. Usually, there are multiple observations at distinct values in the large survey data set. Large-scale data sets not only can mask informative trends between the response variable and the covariates when plotting the data, they also make the estimation process very computationally cumbersome. Hence, it is then very natural in the complex survey data analysis to bin the data into domains according to the distinct values of the characteristic variables. The idea of binning the survey data by the distinct values originates from the scale-load estimators introduced by Hartley and Rao (1968, 1969). A scale-load estimator is computed based on the observations and the frequency within one bin or one domain. A domain mean is a generalization of a scale-load estimator because it is calculated by dividing the weighted sum within each domain by the number of observations within each domain, that is, at each distinct point. To obtain the survey estimator of $m(x)$, we follow the methodology of local polynomial regression in Bellhouse and Stafford (2001), who take advantage of the idea of binning in model-assisted regression analysis.

Suppose there are multiple observations at q distinct values of covariates x. Let x_d denote the d^{th} distinct value or the d^{th} bin and assume that the values of x_d are equally spaced with length $x_d - x_{d-1}$. The population proportion of the observations with x_d is denoted by P_d. Let the vector of population means for response variable y at distinct values of x be $\overline{\mathbf{Y}} = (\overline{Y}_1, \ldots, \overline{Y}_q)$. We denote $\overline{y}_d$ and p_d as the survey estimators of $\overline{Y}_d$ and P_d, respectively. Notice that $\overline{y}_d$ and p_d can be considered as estimators of domain means.

Hence, based on the working model 1, we have the survey estimates of $m(x)$ at x,

$$\hat{m}(x) = \mathbf{e}^T (\mathbf{X}_x^T \hat{\mathbf{KW}} \mathbf{X}_x)^{-1} \mathbf{X}_x^T \hat{\mathbf{KW}} \overline{\mathbf{y}} \tag{5}$$

where

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^l \\ 1 & x_2 - x & \cdots & (x_2 - x)^l \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_q - x & \cdots & (x_q - x)^l \end{bmatrix}$$

and

$$\hat{\mathbf{KW}} = \frac{1}{h} diag\left(p_1 K\left(\frac{x_1 - x}{h}\right), \ldots, p_q K\left(\frac{x_q - x}{h}\right)\right)$$

where $K(\cdot)$ is a kernel function satisfying $\int K(t)dt = 1$ and $\int K(t)^2\, dt < \infty$ and l is the degree of local polynomial the model fits. h is the bandwidth that controls the size of neighbourhood. The vector $\mathbf{e}$ is the $q \times 1$ vector of the form $(1,0, \ldots, 0)^T$. Given that (a, x) is the interval from the lower bound of the covariate x to the point where we evaluate the shift function and based on the above survey estimators, we have the survey estimates of the shift function as $\hat{\Delta}(x) = d(\hat{\mathbf{B}}, \hat{T}_{\mathbf{S}}(x)) - x$, which can be expanded with Taylor series technique as follows,

$$\hat{\Delta}(x) \cong \Delta(x) + \mathbf{d}_{\mathbf{B}}^T (\hat{\mathbf{B}} - \mathbf{B}) + d_T (\hat{T}_{\mathbf{S}}(x) - T_{\mathbf{S}}(x)), \tag{6}$$

where $\mathbf{d_B}$ and d_T are the first derivatives of $\hat{\Delta}(x)$ with respect to $\hat{\mathbf{B}}$ and $\hat{T}_{\mathbf{S}}(x)$, respectively. Hence, the variance of $\hat{\Delta}(x)$ can be obtained from the variances of $\hat{\mathbf{B}}$ and $\hat{T}_{\mathbf{S}}(x)$ and the covariance between $\hat{\mathbf{B}}$ and $\hat{T}_{\mathbf{S}}(x)$. The variance is in the form,

$$V_p(\hat{\Delta}(x)) \cong \mathbf{d}_{\mathbf{B}}^T V_p(\hat{\mathbf{B}}) \mathbf{d}_{\mathbf{B}} + d_T^2 V_p(\hat{T}_{\mathbf{S}}(x)) + 2 d_t \mathbf{d}_{\mathbf{B}}^T Cov_p(\hat{\mathbf{B}}, \hat{T}_{\mathbf{S}}(x)). \tag{7}$$

Given $\hat{m}(t_i)$ is the estimate of $m(t_i)$ for t_i □(a, x), $\hat{T}_{\mathbf{S}}(x) = \sum_{i \in (a,x)} \hat{m}(t_i)$. From (5) we have

$$\hat{T}_{\mathbf{S}}(x) \cong \mathbf{S}(x) \overline{\mathbf{y}}, \tag{8}$$

where $\mathbf{S}(x) = \sum_{t_i \in (a,x)} \mathbf{e}^T (\mathbf{X}_x^T \mathbf{KW} \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{KW}$. Hence, the design-based variance of $\hat{T}_{\mathbf{S}}(x)$ is,

$$V_p(\hat{T}_{\mathbf{S}}(x)) = \mathbf{S}(x) V_p(\overline{\mathbf{y}}) \mathbf{S}(x)^T. \tag{9}$$

Denote vector $\hat{\mathbf{u}}(\mathbf{B}) = \sum_{k \in \mathbf{s}} (y_k - \mathbf{x}_k^T \mathbf{B}) \mathbf{x}_k w_k$ as sum of the products between the parametric regression residuals from the parametric fit and the k^{th} row of $\mathbf{X}(x)$ under the null hypothesis. Based on Binder (1983), we find the variance of $\hat{\mathbf{B}}$ is,

$$V_p(\hat{\mathbf{B}}) \cong (\mathbf{X}(x)^T \mathbf{X}(x))^{-1} \Sigma(\hat{\mathbf{B}}) ((\mathbf{X}(x)^T \mathbf{X}(x))^{-1})^T. \tag{10}$$

where $\Sigma(\hat{\mathbf{B}})$ is the variance-covariance matrix of $\hat{\mathbf{u}}(\mathbf{B})$. Further, based on the Taylor's linearization technique in Binder (1983), we have

$$\hat{\mathbf{B}} - \mathbf{B} \cong -(\mathbf{X}(x)^T \mathbf{X}(x))^{-1} \hat{\mathbf{u}}(\mathbf{B}). \tag{11}$$

Hence, it can be shown that,

$$Cov_p(\hat{\mathbf{B}}, \hat{T}_{\mathbf{s}}(x)) \cong -(\mathbf{X}(x)^T \mathbf{X}(x))^{-1} Cov_p(\hat{\mathbf{u}}(\mathbf{B}), \overline{\mathbf{y}}) \mathbf{S}(x)^T. \tag{12}$$

Let us consider only the j^{th} element of the vector $\hat{\mathbf{u}}(\mathbf{B})$ and write it as $\hat{u}_j(\mathbf{B}) = \sum_{k \in \mathbf{s}} u_k$. Using the idea of binning and denoting $\mathbf{s}_d$ as the set of sample units that fall in the d^{th} domain, we can rewrite $\hat{u}_j(\mathbf{B}) = \sum_{d=1}^{q} \sum_{k \in \mathbf{s}_d} u_k$, or $\hat{u}_j(\mathbf{B}) = \mathbf{l}\hat{\mathbf{t}}_{\mathbf{j}}$, where $\mathbf{l}$ is a 1×q vector with 1's being the entries and $\hat{\mathbf{t}}_{\mathbf{j}}$ is a q×1 vector of the binned totals of the forms, $\sum_{k \in \mathbf{s}_d} u_k$, for $d = 1, \ldots, q$. Hence, by defining

$$\hat{\mathbf{u}}(\mathbf{B}) = (\mathbf{l}\hat{\mathbf{t}}_1, \ldots, \mathbf{l}\hat{\mathbf{t}}_p) \tag{13}$$

we have $Cov_p(\hat{\mathbf{u}}(\mathbf{B}), \overline{\mathbf{y}}) = \left(\mathbf{l} Cov_1(\hat{\mathbf{t}}_1, \overline{\mathbf{y}}), \ldots \mathbf{l} Cov_p(\hat{\mathbf{t}}_p, \overline{\mathbf{y}})\right)^T$. Finally, if we replace all the variances and covariance terms in (7) with the results in (9), (10) and (12), we have the variance of the survey estimates of the shift function, $V_p(\hat{\Delta}(x))$. The estimate of $V_p(\hat{\Delta}(x))$ is then,

$$\hat{V}_p(\hat{\Delta}(x)) \cong \hat{\mathbf{d}}_{\mathbf{B}}^T \hat{V}_p(\hat{\mathbf{B}})\hat{\mathbf{d}}_{\mathbf{B}} + \hat{d}_{\tau}^2 \hat{V}_p(\hat{T}_{\mathbf{S}}(x)) + 2\hat{d}_t \hat{\mathbf{d}}_{\mathbf{B}}^T C\hat{o}v_p(\hat{\mathbf{B}}, \hat{T}_{\mathbf{S}}(x)),$$

where $\hat{\mathbf{d}}_{\mathbf{B}}^T$, $\hat{d}_{\tau}^2$, $\hat{V}_p(\hat{\mathbf{B}})$, $\hat{V}_p(\hat{T}_{\mathbf{S}}(x))$ and $C\hat{o}v_p(\hat{\mathbf{B}}, \hat{T}_{\mathbf{S}}(x))$ are survey estimates of $\mathbf{d}_{\mathbf{B}}^T$, d_{τ}^2, $V_p(\hat{\mathbf{B}})$, $V_p(\hat{T}_{\mathbf{S}}(x))$ and $Cov_p(\hat{\mathbf{B}}, \hat{T}_{\mathbf{S}}(x))$, respectively.

In addition, linearization in (6) suggests that asymptotic normality of $\hat{\Delta}(x)$ is dependent on the properties of $\sqrt{n_{\mathrm{v}}}(\hat{\mathbf{B}} - \mathbf{B})$ and $\hat{T}_{\mathbf{S}}(x)$. Using (9), we can rewrite the Taylor expansion in (6) as,

$$\sqrt{n_{\mathrm{v}}}(\hat{\Delta}(x) - \Delta(x)) \cong -\mathbf{d}_{\mathbf{B}}^T \left(\frac{(\mathbf{X}(x)^T \mathbf{X}(x))}{N_{\mathrm{v}}} \right)^{-1} \frac{\sqrt{n_{\mathrm{v}}}}{N_{\mathrm{v}}} (\hat{\mathbf{u}}(\mathbf{B}) - \mathbf{u}(\mathbf{B})) + d_T \sqrt{n_{\mathrm{v}}}(\hat{T}_{\mathbf{S}}(x) - T_{\mathbf{S}}(x)) \cdot$$

Using the conditions in Binder (1983), we know that $(\mathbf{X}(x)^T \mathbf{X}(x)/N_{\mathrm{v}})^{-1}$ converges to a constant matrix. Additionally, given the identities in (8) and (13), we have in limit,

$$\sqrt{n_{\mathrm{v}}}(\hat{\Delta}(x) - \Delta(x)) \cong \frac{\sqrt{n_{\mathrm{v}}}}{N_{\mathrm{v}}} \sum_{j=1}^{p} c_j \mathbf{l}(\hat{\mathbf{t}}_j - \mathbf{t}_j) + d_T \mathbf{S}(x) \sqrt{n_{\mathrm{v}}}(\bar{\mathbf{y}} - \bar{\mathbf{Y}}), \tag{14}$$

where c_j is the j^{th} element of vector $\mathbf{c} = -\mathbf{d}_{\mathbf{B}}^T \left((\mathbf{X}(x)^T \mathbf{X}(x))/N_{\mathrm{v}} \right)^{-1}$. Since we bin both y and $u_j(\mathbf{B})$ with the same covariate x, we can combine the two terms in (14) and treat it as linear combination of estimated domain means. Based on asymptotic results in Wang and Bellhouse (2004), we can state the asymptotic normality of $\hat{\Delta}(x)$ in the following theorem.

Theorem. *Suppose the shift function is estimated by least squares and a local polynomial regression technique satisfying □n and δ, □Q_0 such that □$q \geq Q^0$,* $\left| \sqrt{n_{\mathrm{v}}} \left(\sum_{i=0}^{q} m(x_i) - T(x) \right) \right| \leq \delta / \sqrt{n_{\mathrm{v}}}$ *. Then, when the null hypothesis is true, we have as v □ ∞,* $\sqrt{n_{\mathrm{v}}}(\hat{\Delta}(x) - \Delta(x)) \Big/ \sqrt{V_p(\hat{\Delta}(x))} \xrightarrow{d} N(0,1)$.

5. EMPIRICAL ILLUSTRATION

As an empirical example, we apply the shift function plot to data from the 1990 Ontario Health Survey to compare the fits of polynomial and local polynomial regression. The Ontario Health Survey was conducted with a stratified two-stage clustered design. The basic strata were the public health units in the province of Ontario. By further dividing the basic strata into the urban and rural areas, at total of 86 strata were obtained. Within each stratum, enumeration areas are the first stage sampling units and within each of the enumeration area, dwellings were randomly selected as second stage units. Households within each dwelling are clusters of observations. The purpose of this survey was to measure the health status of the people of Ontario and to collect data relating to the risk factors of major causes of mortality in Ontario.

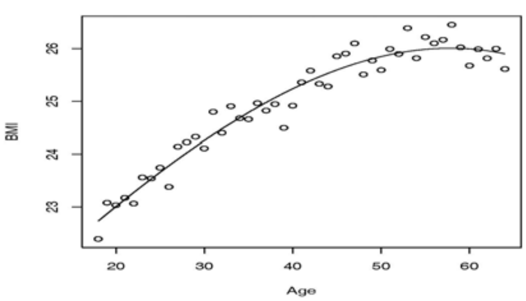

Figure 2: Polynomial Fit of BMI

In this example, we examine the trend of people's weight by age. The measurement we use here is a standardized measure called body mass index (BMI) which is calculated as, BMI = weight in kg/height in meters^2. Because BMI takes the shape of a person's body into account, it is believed to be a better measurement of a person's weight. We repeat one of the examples in Bellhouse and Stafford (2001) and examine age trends in Body Mass Index for people whose ages are between 18 and 64. After fitting the data with both polynomial and local polynomial regression model models as shown in **Figure 2** and **Figure 3** respectively, we plot the estimated shift function to examine the parametric fit, in this case the polynomial of degree 3, i … **e 3**
capture … arame … **ure**
2. Spec … 30 to … cal
polyno … is sub … lly,
the shi … n the … ero
line as …

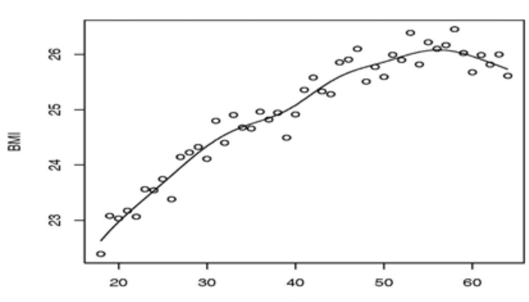

Figure 3: Local Polynomial Fit of BMI

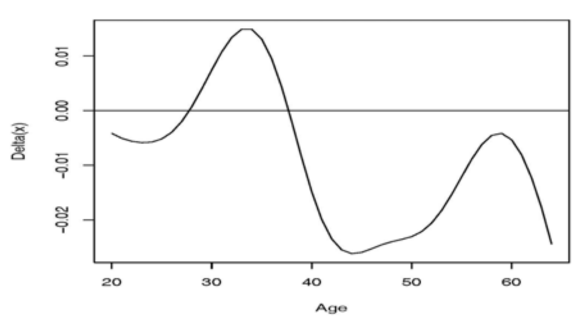

Figure 4: Shift Function Plot for Comparison of Age Trend in BMI

6. CONCLUSION AND FUTURE WORK

In this paper, we developed a new graphic diagnostic tool, shift function plots, to examine the fit of a parametric regression model under the assumption that a nonparametric regression technique can fit a regression model more closely to the true regression model. We considered the construction of shift function plots based on data collected through a complex survey. With the establishment if the working models and the assumption of the nested sequence of finite populations, we developed the asymptotic results in the survey context. In particular, asymptotic normality and the central limit theorem of the estimates of shift function were set up. An empirical example was used to illustrate the application of the estimated shift function with 1990 Ontario Health Survey.

REFERENCES

Bellhouse, D. R. and J. E. Stafford (1997). "Local polynomial regression in complex surveys." *Survey Methodology*, **27**(**2**), 197-203.

Binder, D. A. (1983). "On the variance of asymptotically normal estimators from complex surveys." *International Statistical Review*, **51**, 279-292.

Doksum, K. A. and G. L. Sievers (1976). "Plotting with confidence: graphical comparisons of two populations." *Biometrika*, **63**(**3**), 421-434.

Fuller, W. A. (1975). "Regression analysis for sample surveys." *Sankhyā*, **C37**, 117-132.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*. Spinge.

Hartley, H. O. and J. N. K. Rao (1968). "A new estimation theory for sample surveys." *Biometrika*, **55**(**3**), 547-557.

Hartley, H. O. and J. N. K. Rao (1969). "A new estimation theory for sample surveys: II." In N. L. Johnson and H. Smith (Eds.) *New Developments in Survey Sampling*, 147-169.

Wang, Z. (2004). *Some Nonparametric Regression Techniques for Complex Surveys*. Unpublished Ph.D. thesis, University of Western Ontario.

Wang, Z. and D. R. Bellhouse (2004). Asymptotic Properties of Domain Means. Submitted for publication.