On Imputation Techniques in Survey Sampling

Hui Rong Zhu

A Thesis
in
the Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirement
for the Degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

August 2014

Concordia University

School of Graduate Studies

This is to vertify that the thesis prepared

By: HuiRong Zhu

Entitled: On Imputation Techniques in Survey Sampling

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Mathematics and Statistics)

complies with the regulations of the University and meets the accepted standard with respect to orginality and quality.

Signed by the final Examining committee:

_____ Examiner

Dr. A.Sen

_____ Examiner

Dr. D.Sen

_____ Supervisor

Dr. Y P. Chaubey

Approved by

_____

Dr J.Garrido(Graduate Program Director) Date

_____

Dean of Facault of Arts and Science Date

Abstract

On Imputation Techniques in Survey Sampling

Hui Rong Zhu

Some nonparametric imputation techniques, including two categories: single imputation and multiple imputation, are introduced and studied. Some properties of the estimators such as the bias, the variance, and the mean squared error are presented. Finally, some imputation techniques are applied to a real case. These methods are compared in order to assess their advantages, disadvantages, and applicabilities.

Acknowledgment

I express my deepest gratitude to Dr. Y. P. Chaubey for accepting me as his student and for his guidance, advice and great support in the preparation of this thesis. His strong knowledge on statistics has guided me throughout the study.

I also thank Dr. A. Sen and Dr. D. Sen for accepting to work on my thesis committee. As my statistics teachers throughout my bachelor/master study in Concordia University, their excellent lectures allow me to build strong confidence on statistical analysis.

I thank Ms. Marie-France Leclere for her excellent assistance on Administration during my master study and Dr. W.Sun, Dr. E. G. Cohen to give me a help for my graduate study application.

Furthermore, I dedicate this thesis to my mother, FengLan Lu, my father, XiaoLiang Zhu, and my sister, HuiQin Zhu, for their continuous support and love.

Finally, this thesis is gratefully dedicated to my husband Gang Wu and my son Felix Wu for the love. As a mathematician, my husband gave me invaluable reference and encouragement. Without his direction, my study on mathematics/statistics would not be successful. My son is smart and sportive. Being with him, everyday is sunshine for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In sample surveys, missing responses occur frequently,resulting in incomplete sample. These incomplete samples are called as missing data. Missing data may be caused by sensitive questions, improper data collection and so on. If the incomplete data occupy only a small portion of the dataset, the data deletion may be a good way to the missing-data problem. However, in most cases, if we ignore these missing data during the statistical analysis, the results may not be representative. In order to form a complete dataset for the standard analysis, imputation is introduced and it has become one of the most popular techniques used to resolve missing data problems in sampling survey data analyses. Imputation is to replace missing data with a plausible value based on other available informations.

1.1 Mechanisms for Missing Data

Little and Rubin in [17] defined three classes of missing data. These three general missing mechanisms are presented here with examples.

- Missing Completely at Random(MCAR)
  The missing data occurs randomly and doesn't depend on both of observed data and unobserved data. In a sample survey setting, MCAR is sometimes called uniform non-response. For example, If a laboratory sample is dropped, the resulting observation is missing. We can say this is MCAR.

- Missing at Random (MAR)
  Given the observed data, the missingness is not related to the unobserved data. For example, in a survey of relation between property tax band and income, usually these people with higher salary and lower salary may omit to answer the income questions. So given the property tax band, non-response to the income questions is random.

- Observed at Random(OAR)
  Given the observed and unobserved data, the missingness is not related to the observed data. For example, in a survey to examine the effect of education on income, these non-response to income is not OAR if income is a related to education.

Apart from these, there are some other missing mechanisms.

- Missing Not at Random (MNAR)
  If the data is neither MCAR nor MAR, we can say the data is MNAR.

- Not Missing at Random (NMAR)
  The data is missing due to the particular reason.

1.2 Imputation Methods

In general, the imputation methods are divided into two categories: Model-based Imputation and Nonparametric Imputation. The simple imputation methods and the multiple imputation methods are included in these two categories.

In [26], Rubin compares both of the single imputation and multiple imputation methods. He comments on the of these advantage and disadvantage as follow.

i) Simple Imputation: This type of method replaces the missing data once by a randomly selected response value.

- Advantage:

- The standard complete-data methods of statistical analysis can be used if the missing values have been imputed.

- Data collector's knowledge can be incorporated.

- Disadvantage:

- The inference based on the imputed data set may be too sharp as the extra variability due to the unknown missing values is not being taken into account.

ii) Multiple Imputation

Multiple imputation is a statistically principled and commonly used method. The idea of multiple imputation is to repeat the process of assigning several (say m between 2 to 10) values for each missing data. The m imputations for each missing data will create m sets of complete data. Hence, the standard complete-data analysis is conducted for each completed data sets.

- Advantage:

- Multiple imputation increases the efficiency of estimation.

- Disadvantage:

- More work and space are needed to analyze a multiply-imputed data set.

1.3 Imputation Problems

An imputation technique might cause its own problems. In [27], Sande listed out some general problems such as:

- Since the imputed value of the field has to satisfy the reasonable constraints which is known as edits to ensure that the completed data is consistent as well as it will reduce the applicability of the imputation procedure.

- It is hard to determine whether the method of imputation is specified properly and precisely.

- Imputation does not solve the specific problems of estimation better than the tradition estimation techniques for missing data.

The evaluation of imputation technique is, in general, to compare the bias, the variance, and the mean squared error of those estimators.

In this thesis, we focus on the nonparametric imputation methods.

In Chapter 2, some linear single and multiple imputation techniques will be introduced.

In Chapter 3, the kernel smoothing techniques will be reviewed for the nonparametric imputations.

Finally, we have given an application to a real sampling case

in Chapter 4 for some imputation techniques presented in the previous chapters. Then we compare these imputation techniques with their applications.

Chapter 2

Linear Imputation Approaches

2.1 Single Imputation Approaches

There are several methods to handle the problem of incompleteness or nonresponse in a census or a sample survey. One kind of them is called Hot-Deck imputation. Hot-Deck imputation is a common technique to deal with missing data in survey sampling. The major idea of Hot-Deck imputation is to replace the missing data by observable and measurable values from a similar group. By this idea, some specific methods are developed. In this chapter, we introduce some approaches of imputation under two-phase sampling described in [30]. That says, one has two sets of sampling data: Y and X, strongly correlated. The set Y is incomplete, i.e., there are some nonresponse data in Y. The set X is complete, i.e., all the elements in X are observed or measured. The approaches under two-phase sampling Y and X are to establish a relation of elements between these two sets

Y, X and then replace the nonresponse data by the relationship.

Consider a sample survey in a finite population of N units: $\Omega = \{1, 2, \cdots, N\}$. Denote $y_i, i \in \Omega$ the outcome statistical variable that gives a characteristic of the individual i. In order to estimate the mean $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, one draws a random sample without replacement of n units $S = \{1, 2, \cdots, n\} \subseteq \Omega$ from this population, where the number of the responding units in this sample is r. Denote the set of the outcome variables from this sample as

$$Y_S = \{y_1, \cdots, y_n\} = \{y_k : k \in S\}$$

Defines the response indicator

$$R = \Big(R_1, R_2, \ldots, R_N\Big)$$

that indicates which values are respondent or nonrespondent in the survey, where

$$R_i = \begin{cases} 1 & \text{if } y_i \text{ is respondent} \\ 0 & \text{if } y_i \text{ is nonrespondent.} \end{cases} \tag{2.1}$$

Then, $S = S_R \cup S_{NR}$, where S_R and S_{NR} are the sets of respondent units and nonrespondent units respectively:

$$S_R = \{k \in S : R_k = 1\}$$
$$S_{NR} = \{k \in S : R_k = 0\}$$

In order to estimate the nonrespondent values, one needs another phase sampling data, the covariates $X = \{x_i : i \in \Omega\}$, that describe a characteristic of individuals fully observed or measured. Similarly, denote the set of the covariates from the sample S as

$$X_S = \{x_1, \cdots, x_n\} = \{x_k : k \in S\}$$

For the nonrespondent units $\{k \in S : R_k = 0\}$, one assumes that y_k is a function of X_S:

$$y_k = h_k(X_S), \text{if } R_k = 0 \tag{2.2}$$

Then, one makes the imputation:

$$y_{Ik} = \begin{cases} y_k & \text{if } R_k = 1 \\ h_k(X_S) & \text{if } R_k = 0 \end{cases} \tag{2.3}$$

The estimation of $\bar{y}$ is given by

$$\bar{y}_{imp} = \frac{1}{n} \sum_{k=1}^{n} y_{Ik} \tag{2.4}$$

For using this imputation, one defines some means from the

sample and the responding values:

$$\bar{x}_n = \frac{1}{n}\sum_{k\in S} x_k$$
$$\bar{x}_r = \frac{1}{r}\sum_{k\in S_R} x_k$$
$$\bar{y}_r = \frac{1}{r}\sum_{k\in S_R} y_k$$

and the estimators of variance and the covariance:

$$\begin{aligned} s_{xy} &= \frac{1}{r-1}\sum_{k=1}^{r}(x_k - \bar{x}_r)(y_k - \bar{y}_r) \\ s_x^2 &= \frac{1}{r-1}\sum_{k=1}^{r}(x_k - \bar{x}_r)^2 \\ s_y^2 &= \frac{1}{r-1}\sum_{k=1}^{r}(y_k - \bar{y}_r)^2 \end{aligned} \tag{2.5}$$

where

$$r = \sum_{i=1}^{n} R_i$$

In practice, the relation in (2.2) may be assumed linear:

$$h(x_k) = A + Bx_k, \qquad \text{if } R_k = 0 \tag{2.6}$$

Then, the imputation becomes

$$y_{Ik} = \begin{cases} y_k & \text{if } R_k = 1 \\ A + Bx_k & \text{if } R_k = 0 \end{cases} \tag{2.7}$$

To analyze the errors, one defines

$$\epsilon = \frac{\bar{y}_r}{\bar{y}} - 1, \qquad \delta = \frac{\bar{x}_r}{\bar{x}} - 1, \qquad \eta = \frac{\bar{x}_n}{\bar{x}} - 1$$

where,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Suppose that the expectations $E(x_k) = \mu_X = \bar{x}$ and $E(y_k) = \mu_Y = \bar{y}$ for every $k \in \Omega$. Then, one sees immediately that $E(\bar{y}_r) = \bar{y}$, $E(\bar{x}_r) = E(\bar{x}_n) = \bar{x}$, and thus,

$$E(\epsilon) = 0, \qquad E(\delta) = 0, \qquad E(\eta) = 0 \tag{2.8}$$

and one has

$$\begin{aligned} E(\epsilon^2) &= E((\frac{\bar{y}_r}{\bar{y}} - 1)^2) = \frac{1}{\bar{y}^2}E((\bar{y}_r - \bar{y})^2) = \frac{1}{\bar{y}^2}E(\bar{y}_r^2 - 2\bar{y}_r\bar{y} + \bar{y}^2) \\ &= \frac{1}{\bar{y}^2}[E(\bar{y}_r^2) - 2E(\frac{1}{r}\sum_{i=1}^{r} y_i)\bar{y} + \bar{y}^2] = \frac{1}{\bar{y}^2}[E(\bar{y}_r^2) - \bar{y}^2] \\ &= \frac{1}{\bar{y}^2}[Var(\bar{y}_r) + (E(\bar{y}_r))^2 - \bar{y}^2] = \frac{1}{\bar{y}^2}Var(\bar{y}_r) \\ &= \frac{1}{\bar{y}^2}S^2_{\bar{y}_r} = \frac{s_y^2}{\bar{y}^2 r}(1 - \frac{r}{N}). \end{aligned} \tag{2.9}$$

Similarly,

$$E(\delta^2) = E((\frac{\bar{x}_r}{\bar{x}} - 1)^2) \ = \frac{1}{\bar{x}^2}E((\bar{x}_r - \bar{x})^2) = \frac{1}{\bar{x}^2}E(\bar{x}_r^2 - 2\bar{x}_r\bar{x} + \bar{x}^2)$$
$$= \frac{1}{\bar{x}^2}[E(\bar{x}_r^2) - 2E(\frac{1}{r}\sum_{i=1}^{r} x_i)\bar{x} + \bar{x}^2] = \frac{1}{\bar{x}^2}[E(\bar{x}_r^2) - \bar{x}^2]$$
$$= \frac{1}{\bar{x}^2}[Var(\bar{x}_r) + (E(\bar{x}_r))^2 - \bar{x}^2] = \frac{1}{\bar{x}^2}Var(\bar{x}_r)$$
$$= \frac{1}{\bar{x}^2}S^2_{\bar{x}_r} = \frac{s_x^2}{\bar{x}^2 r}(1 - \frac{r}{N}) \tag{2.10}$$

$$E(\eta^2) = \frac{S_x^2}{\bar{x}^2 n}(1 - \frac{n}{N}) \approx \frac{s_x^2}{\bar{x}^2 n}(1 - \frac{n}{N}) \tag{2.11}$$

where

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

and one has

$$E(\epsilon\eta) = E((\frac{\bar{y}_r}{\bar{y}} - 1)(\frac{\bar{x}_n}{\bar{x}} - 1)) = \frac{1}{\bar{y}\bar{x}}E(\bar{y}_r\bar{x}_n - \bar{y}_r\bar{x} - \bar{y}\bar{x}_n + \bar{y}\bar{x})$$
$$= \frac{1}{\bar{y}\bar{x}}E[(\bar{y}_r(\bar{x}_n - \bar{x}) - \bar{y}(\bar{x}_n - \bar{x})]$$
$$= \frac{1}{\bar{y}\bar{x}}E[(\bar{y}_r - \bar{y})(\bar{x}_n - \bar{x})]$$
$$= \frac{1}{\bar{y}\bar{x}}Cov(\bar{y}_r, \bar{x}_n)$$
$$= \frac{1}{\bar{y}\bar{x}}s_{\bar{y}_r\bar{x}_n} \approx \frac{s_{xy}}{n\bar{y}\bar{x}}(1 - \frac{n}{N}). \tag{2.12}$$

Similarly,

$$E(\epsilon\delta) = \frac{1}{\bar{y}\bar{x}}Cov(\bar{y}_r, \bar{x}_r) = \frac{1}{\bar{y}\bar{x}}s_{\bar{y}_r\bar{x}_r} \approx \frac{s_{xy}}{r\bar{y}\bar{x}}(1 - \frac{r}{N}) \tag{2.13}$$

$$E(\delta\eta) = \frac{1}{\bar{x}^2}Cov(\bar{x}_r, \bar{x}_n) = \frac{1}{\bar{x}^2}s_{\bar{x}_r\bar{x}_n} \approx \frac{s_x^2}{n\bar{x}^2}(1 - \frac{n}{N}) \tag{2.14}$$

2.1.1 Mean Method of Imputation

This is the simplest method of imputation. All missing values y_k are just replaced by the mean of responding values $\bar{y}_r$, i.e., $A = \bar{y}_r$ and $B = 0$ in (2.6). Thus (2.7) gives

$$y_{Ik} = \begin{cases} y_k & \text{if } R_k = 1 \\ \bar{y}_r & \text{if } R_k = 0 \end{cases} \tag{2.15}$$

This method effectively ignores all nonresponse data and simply represents the nonresponse data by the mean of responding data, as (2.4), in this case, becomes

$$\begin{aligned} \bar{y}_{mean} = \bar{y}_{imp} &= \frac{1}{n}\sum_{k=1}^{n} y_{Ik} \\ &= \frac{1}{n}(\sum_{i=1}^{r} y_i + (n-r)\bar{y}_r) = \frac{1}{r}\sum_{i=1}^{r} y_i = \bar{y}_r \end{aligned} \tag{2.16}$$

In a survey, the nonresponse data might have a different view from the responding data for some specific reasons. Consequently, such representation would not be correct. It risks to lose or distort the true image. This implies that the mean method could not resolve the imputation that problems by nonresponse that we mentioned in Chapter 1.

The estimator (2.16) can be written in terms of ϵ:

$$\bar{y}_{mean} = \frac{\bar{y}_r}{\bar{y}}\bar{y} = \bar{y}(1+\epsilon) \tag{2.17}$$

Thus, the variance of $\bar{y}_{mean}$ is given by

$$\begin{aligned} Var(\bar{y}_{mean}) &= Var(\bar{y}(1+\epsilon)) = \bar{y}^2 Var(\epsilon) = \bar{y}^2[E(\epsilon^2) - (E(\epsilon))^2] \\ &= \bar{y}^2 E(\epsilon^2) = \bar{y}^2 \frac{s_y^2}{\bar{y}^2 r}(1 - \frac{r}{N}) = \frac{s_y^2}{r}(1 - \frac{r}{N}) \end{aligned} \tag{2.18}$$

and the mean squared error of $\bar{y}_{mean}$ is given by

$$\begin{aligned} MSE(\bar{y}_{mean}) &= E[(\bar{y}_{mean} - \bar{y})^2] = \bar{y}^2 E[(\frac{\bar{y}_r - \bar{y}}{\bar{y}})^2] \\ &= \bar{y}^2 E(\epsilon^2) = \bar{y}^2 \frac{s_y^2}{\bar{y}^2 r}(1 - \frac{r}{N}) = \frac{s_y^2}{r}(1 - \frac{r}{N}). \end{aligned} \tag{2.19}$$

2.1.2 Ratio Method of Imputation

This method improves the mean method by introducing a ratio $\frac{x_k}{\bar{x}_r}$ for every missing unit $k \in S_{NR}$. The missing values y_k are replaced by the ratio of the mean of responding values: $y_k =$

$\frac{x_k}{\bar{x}_r}\bar{y}_r$, i.e., $A = 0$ and $B = \frac{\bar{y}_r}{\bar{x}_r}$ in (2.6). Thus (2.7) gives

$$y_{Ik} = \begin{cases} y_k & \text{if } R_k = 1 \\ \frac{\bar{y}_r}{\bar{x}_r} x_k & \text{if } R_k = 0 \end{cases} \tag{2.20}$$

The estimator (2.4) becomes

$$\begin{aligned} \bar{y}_r = \bar{y}_{imp} &= \frac{1}{n}\sum_{k=1}^{n} y_{Ik} \\ &= \frac{1}{n}(\sum_{i=1}^{r} y_i + \frac{\bar{y}_r}{\bar{x}_r}\sum_{i=r+1}^{n} x_i) = \frac{\bar{y}_r}{n\bar{x}_r}(r\bar{x}_r + \sum_{i=r+1}^{n} x_i) \\ &= \frac{\bar{y}_r}{n\bar{x}_r}\sum_{k=1}^{n} x_k = \bar{y}_r(\frac{\bar{x}_n}{\bar{x}_r}) \end{aligned} \tag{2.21}$$

The estimator (2.21) can be written in terms of ϵ, δ, and η:

$$\begin{aligned} \bar{y}_r &= \frac{\bar{y}_r}{\bar{y}}\bar{y}(\frac{\bar{x}_n/\bar{x}}{\bar{x}_r/\bar{x}}) = (1+\epsilon)\frac{1+\eta}{1+\delta}\bar{y} = \bar{y}(1+\epsilon)(1+\eta)[\sum_{i=0}^{\infty}(-1)^i\delta^i] \\ &= \bar{y}(1+\epsilon)(1+\eta)(1-\delta+\delta^2+\cdots) = \bar{y}(1+\epsilon+\eta+\epsilon\eta)(1-\delta+\delta^2+\cdots) \\ &= \bar{y}[1+\epsilon+\eta-\delta+\delta^2+\epsilon\eta-\epsilon\delta-\delta\eta+O(\epsilon\eta\delta)] \end{aligned} \tag{2.22}$$

Thus, the variance of $\bar{y}_{ratio}$ is given by

$$\begin{aligned}
Var(\bar{y}_r) &= Var\{\bar{y}[1+\epsilon+\eta-\delta+\delta^2+\epsilon\eta-\epsilon\delta-\delta\eta+O(\epsilon\eta\delta)]\} \\
&= \bar{y}^2[Var(1+\epsilon+\eta-\delta+\delta^2+\epsilon\eta-\epsilon\delta-\delta\eta+O(\epsilon\eta\delta))] \\
&\approx \bar{y}^2[Var(\epsilon)+Var(\eta)+Var(\delta)] \\
&= \bar{y}^2[E(\epsilon^2)-(E(\epsilon))^2+E(\eta^2)-(E(\eta))^2+E(\delta^2)-(E(\delta))^2] \\
&= \bar{y}^2[E(\epsilon^2)+E(\eta^2)+E(\delta^2)]
\end{aligned} \tag{2.23}$$

and the mean squared error of $\bar{y}_{ratio}$ to the first order of approximation is given by

$$\begin{aligned}
MSE(\bar{y}_r) &= E[(\bar{y}_{ratio}-\bar{y})^2] \\
&= \bar{y}^2E[(\epsilon+\eta-\delta+\delta^2+\epsilon\eta-\epsilon\delta-\delta\eta+O(\epsilon\eta\delta))^2] \\
&\approx \bar{y}^2E[(\epsilon+\eta-\delta)^2] \\
&= \bar{y}^2E(\epsilon^2+\eta^2+\delta^2+2\epsilon\eta-2\epsilon\delta-2\delta\eta) \\
&\approx (\frac{1}{r}-\frac{1}{N})s_y^2+\bar{y}^2(\frac{1}{n}-\frac{1}{N})\frac{s_x^2}{\bar{x}^2}+\bar{y}^2(\frac{1}{r}-\frac{1}{N})\frac{s_x^2}{\bar{x}^2} \\
&\quad +2\bar{y}^2(\frac{1}{n}-\frac{1}{N})\frac{s_{xy}}{\bar{y}\bar{x}^2}-2\bar{y}^2(\frac{1}{r}-\frac{1}{N})\frac{s_{xy}^2}{\bar{y}\bar{x}^2}-2\bar{y}^2(\frac{1}{n}-\frac{1}{N})\frac{s_x^2}{\bar{x}^2} \\
&= (\frac{1}{r}-\frac{1}{N})s_y^2+(\frac{1}{r}-\frac{1}{n})(\frac{\bar{y}}{\bar{x}})^2s_x^2-2(\frac{1}{r}-\frac{1}{n})\frac{\bar{y}}{\bar{x}}s_{xy}
\end{aligned} \tag{2.24}$$

Note that if $\frac{\bar{y}}{\bar{x}} < 2\frac{s_{xy}}{s_x^2}$,

$$\begin{aligned} MSE(\bar{y}_r) &= (\frac{1}{r} - \frac{1}{N})s_y^2 + (\frac{1}{r} - \frac{1}{n})(\frac{\bar{y}}{\bar{x}})^2 s_x^2 - 2(\frac{1}{r} - \frac{1}{n})\frac{\bar{y}}{\bar{x}}s_{xy} \\ &< (\frac{1}{r} - \frac{1}{N})s_y^2 + (\frac{1}{r} - \frac{1}{n})(4\frac{s_{xy}^2}{s_x^4}s_x^2 - 4\frac{s_{xy}^2}{s_x^2}) \\ &= (\frac{1}{r} - \frac{1}{N})s_y^2 = MSE(\bar{y}_{mean}) \end{aligned}$$

It follows that the ratio method of imputation is better than the mean method of imputation if

$$\frac{\bar{y}}{\bar{x}} < 2\frac{s_{xy}}{s_x^2} \tag{2.25}$$

2.1.3 Regression Method of Imputation

In (2.6), choosing $A = \bar{y}_r - \frac{s_{xy}}{s_x^2}\bar{x}_r$ and $B = \frac{s_{xy}}{s_x^2}$, one obtains the regression method of imputation:

$$y_{Ik} = \begin{cases} y_k & \text{if } R_k = 1 \\ \bar{y}_r - \frac{s_{xy}}{s_x^2}\bar{x}_r + \frac{s_{xy}}{s_x^2}x_k & \text{if } R_k = 0 \end{cases} \tag{2.26}$$

The estimator (2.4) becomes

$$\begin{aligned}
\bar{y}_{reg} = \bar{y}_{imp} &= \frac{1}{n}\sum_{k=1}^{n} y_{Ik} \\
&= \frac{1}{n}[\sum_{k=1}^{r} y_k + \sum_{k=r+1}^{n} (\bar{y}_r - \frac{s_{xy}}{s_x^2}\bar{x}_r + \frac{s_{xy}}{s_x^2}x_k)] \\
&= \frac{1}{n}[r\bar{y}_r + (n-r)(\bar{y}_r - \frac{s_{xy}}{rs_x^2}\sum_{k=1}^{r} x_k) + \frac{s_{xy}}{s_x^2}\sum_{k=r+1}^{n} x_k] \\
&= \bar{y}_r - \frac{s_{xy}}{rs_x^2}\sum_{k=1}^{r} x_k + \frac{s_{xy}}{ns_x^2}(\sum_{k=1}^{r} x_k + \sum_{k=r+1}^{n} x_k) \\
&= \bar{y}_r - \frac{s_{xy}}{s_x^2}\bar{x}_r + \frac{s_{xy}}{s_x^2}\bar{x}_n
\end{aligned} \tag{2.27}$$

The estimator (2.27) can be written in terms of ϵ, δ, and η:

$$\begin{aligned}
\bar{y}_{reg} &= \frac{\bar{y}_r}{\bar{y}}\bar{y} + \frac{s_{xy}}{s_x^2}(\frac{\bar{x}_n}{\bar{x}} - \frac{\bar{x}_r}{\bar{x}})\bar{x} \\
&= \bar{y}(1+\epsilon) + \bar{x}\frac{s_{xy}}{s_x^2}(\eta - \delta)
\end{aligned} \tag{2.28}$$

Thus, the mean squared error of $\bar{y}_{regression}$ is given by

$$\begin{aligned}
MSE(\bar{y}_{reg}) &= E[(\bar{y}_{regression} - \bar{y})^2] \\
&= E[(\bar{y}\epsilon + \bar{x}\frac{s_{xy}}{s_x^2}(\eta - \delta))^2] \\
&= E[\bar{y}^2\epsilon^2 + 2\bar{y}\bar{x}\frac{s_{xy}}{s_x^2}\epsilon(\eta - \delta) + \bar{x}^2(\frac{s_{xy}}{s_x^2})^2(\eta - \delta)^2] \\
&= \bar{y}^2E(\epsilon^2) + 2\bar{y}\bar{x}\frac{s_{xy}}{s_x^2}[E(\epsilon\eta) - E(\epsilon\delta)] \\
&\quad + \bar{x}^2(\frac{s_{xy}}{s_x^2})^2[E(\eta^2) - 2E(\eta\delta) + E(\delta^2)] \\
&\approx \bar{y}^2(\frac{1}{r} - \frac{1}{N})\frac{s_y^2}{\bar{y}^2} + 2\bar{y}\bar{x}\frac{s_{xy}}{s_x^2}[(\frac{1}{n} - \frac{1}{N})\frac{s_{xy}}{\bar{x}\bar{y}} - (\frac{1}{r} - \frac{1}{N})\frac{s_{xy}}{\bar{x}\bar{y}}] \\
&\quad + \bar{x}^2(\frac{s_{xy}}{s_x^2})^2[(\frac{1}{n} - \frac{1}{N})\frac{s_x^2}{\bar{x}^2} - 2(\frac{1}{n} - \frac{1}{N})\frac{s_x^2}{\bar{x}^2} + (\frac{1}{r} - \frac{1}{N})\frac{s_x^2}{\bar{x}^2}] \\
&= (\frac{1}{r} - \frac{1}{N})s_y^2 - (\frac{1}{r} - \frac{1}{n})\frac{s_{xy}^2}{\bar{x}^2}
\end{aligned} \tag{2.29}$$

Comparing (2.29) with (2.24) and (2.19), one sees that the regression method of imputation is better than the mean and ratio methods of imputation.

2.1.4 Power Transformation Method of Imputation

In (2.6), choosing $A = 0$ and $B = \bar{y}_r\frac{n(\frac{\bar{x}_n}{\bar{x}_r})^\alpha - r}{n\bar{x}_n - r\bar{x}_r}$, where α is a chosen constant, one obtains the power transformation method

of imputation:

$$y_{Ik} = \begin{cases} y_k & \text{if } R_k = 1 \\ \bar{y}_r \frac{n(\frac{\bar{x}_n}{\bar{x}_r})^\alpha - r}{n\bar{x}_n - r\bar{x}_r} x_k & \text{if } R_k = 0 \end{cases} \tag{2.30}$$

The estimation (2.4) becomes

$$\begin{aligned} \bar{y}_{power} = \bar{y}_{imp} &= \frac{1}{n}\sum_{k=1}^{n} y_{Ik} \\ &= \frac{1}{n}[\sum_{k=1}^{r} y_k + \sum_{k=r+1}^{n} \bar{y}_r \frac{n(\frac{\bar{x}_n}{\bar{x}_r})^\alpha - r}{n\bar{x}_n - r\bar{x}_r} x_k] \\ &= \frac{r\bar{y}_r}{n} + \bar{y}_r \frac{n(\frac{\bar{x}_n}{\bar{x}_r})^\alpha - r}{n\bar{x}_n - r\bar{x}_r} (\frac{n\bar{x}_n - r\bar{x}_r}{n}) \\ &= P\bar{y}_r + B(\bar{x}_n - P\bar{x}_r) \end{aligned} \tag{2.31}$$

where the response rate $P = \frac{r}{n}$

In [29], Singh and Deo declared that the power transformation method is as good as the regression method of imputation.So we have more choice for the imputation.

2.1.5 Optimal Method of Imputation

The optimal method of imputation is to find the coefficients A and B in (2.7) so that the mean squared error of the proposed estimator $\bar{y}_{imp}$ is minimized.

By (2.7), the estimator (2.4) can be written as

$$\begin{aligned}\bar{y}_{imp} &= \frac{1}{n}\sum_{k=1}^{n} y_{Ik} = \frac{1}{n}[\sum_{k=1}^{r} y_k + \sum_{k=r+1}^{n}(A + Bx_k)] \\ &= \frac{1}{n}[r\bar{y}_r + (n-r)A + B(\sum_{k=1}^{r} x_k + \sum_{k=r+1}^{n} x_k) - B\sum_{k=1}^{r} x_k] \\ &= \frac{r}{n}\bar{y}_r + (1-\frac{r}{n})A + B(\bar{x}_n - \frac{r}{n}\bar{x}_r)\end{aligned} \tag{2.32}$$

Noting that $E(\bar{y}_r) = \bar{y}$, $E(\bar{x}_r) = \bar{x}$ and $E(\bar{x}_n) = \bar{x}$, one obtains the bias of the estimator $\bar{y}_{imp}$:

$$\begin{aligned}Bias(\bar{y}_{imp}) &= E(\bar{y}_{imp}) - \bar{y} \\ &= \frac{r}{n}E(\bar{y}_r) + (1-\frac{r}{n})A + B(E(\bar{x}_n) - \frac{r}{n}E(\bar{x}_r)) - \bar{y} \\ &= \frac{r}{n}\bar{y} + (1-\frac{r}{n})A + B\bar{x}(1-\frac{r}{n}) - \bar{y} \\ &= (1-\frac{r}{n})(A + B\bar{x} - \bar{y})\end{aligned} \tag{2.33}$$

This implies that under the assumption $r < n$, the method (2.7) is unbiased if

$$A + B\bar{x} - \bar{y} = 0 \tag{2.34}$$

The estimator (2.32) can be written in terms of ϵ, δ, and η:

$$\begin{aligned}
\bar{y}_{imp} &= \frac{r}{n}\frac{\bar{y}_r}{\bar{y}}\bar{y} + (1-\frac{r}{n})A + B(\frac{\bar{x}_n}{\bar{x}}\bar{x} - \frac{r}{n}\frac{\bar{x}_r}{\bar{x}}\bar{x}) \\
&= \frac{r}{n}(1+\epsilon)\bar{y} + (1-\frac{r}{n})A + B[(1+\eta)\bar{x} - \frac{r}{n}(1+\delta)\bar{x}] \\
&= \frac{r}{n}(1+\epsilon)\bar{y} + (1-\frac{r}{n})A + B\bar{x}(1-\frac{r}{n}+\eta-\frac{r}{n}\delta)
\end{aligned} \tag{2.35}$$

The mean squared error of $\bar{y}_{imp}$ is given by

$$\begin{aligned}
MSE(\bar{y}_{imp}) &= E[(\bar{y}_{imp}-\bar{y})^2] \\
&= E[(\frac{r}{n}(1+\epsilon)\bar{y} + (1-\frac{r}{n})A + B\bar{x}(1-\frac{r}{n}+\eta-\frac{r}{n}\delta) - \bar{y})^2] \\
&= E[(\frac{r}{n}\bar{y}\epsilon + B\bar{x}(\eta-\frac{r}{n}\delta) + U)^2] \\
&= E[(\frac{r}{n}\bar{y}\epsilon + B\bar{x}(\eta-\frac{r}{n}\delta))^2 + 2U(\frac{r}{n}\bar{y}\epsilon + B\bar{x}(\eta-\frac{r}{n}\delta)) + U^2] \\
&= E[(\frac{r}{n})^2\bar{y}^2\epsilon^2 + 2B\frac{r}{n}\bar{x}\bar{y}\epsilon(\eta-\frac{r}{n}\delta) + B^2\bar{x}^2(\eta-\frac{r}{n}\delta)^2] \\
&\quad + E[2U(\frac{r}{n}\bar{y}\epsilon + B\bar{x}(\eta-\frac{r}{n}\delta)) + U^2] \\
&= (\frac{r}{n})^2\bar{y}^2E(\epsilon^2) + 2B\frac{r}{n}\bar{x}\bar{y}E(\epsilon\eta) - 2B(\frac{r}{n})^2\bar{x}\bar{y}E(\epsilon\delta) \\
&\quad + B^2\bar{x}^2E(\eta^2) - 2B^2\bar{x}^2\frac{r}{n}E(\eta\delta) + B^2\bar{x}^2(\frac{r}{n})^2E(\delta^2) + U^2
\end{aligned} \tag{2.36}$$

where

$$U = U(A,B) = (1-\frac{r}{n})(A + B\bar{x} - \bar{y})$$

Putting

$$\begin{aligned} 0 &= \frac{\partial MSE(\bar{y}_{imp})}{\partial A} \\ &= 2U\frac{\partial U}{\partial A} = 2(1-\frac{r}{n})^2(A+B\bar{x}-\bar{y}) \end{aligned}$$

and

$$\begin{aligned} 0 &= \frac{\partial MSE(\bar{y}_{imp})}{\partial B} \\ &= 2B\bar{x}^2[E(\eta^2) - \frac{2r}{n}E(\eta\delta) + (\frac{r}{n})^2E(\delta^2)] \\ &\quad + \frac{2r}{n}\bar{x}\bar{y}E(\epsilon\eta) - 2(\frac{r}{n})^2\bar{x}\bar{y}E(\epsilon\delta) + 2U\frac{\partial U}{\partial B} \\ &= 2A(1-\frac{r}{n})^2\bar{x} + 2B\bar{x}^2[E(\eta^2) - \frac{2r}{n}E(\eta\delta) + (\frac{r}{n})^2E(\delta^2)] \\ &\quad + 2B(1-\frac{r}{n})^2\bar{x}^2 + \frac{2r}{n}\bar{x}\bar{y}E(\epsilon\eta) - 2(\frac{r}{n})^2\bar{x}\bar{y}E(\epsilon\delta) - 2(1-\frac{r}{n})^2\bar{x}\bar{y} \end{aligned}$$

one obtains the optimal solution:

$$A = \bar{y} - B\bar{x} \tag{2.37}$$

and

$$B = \frac{(\frac{\bar{y}}{\bar{x}})[(\frac{r}{n})^2E(\epsilon\delta) - \frac{r}{n}E(\epsilon\eta)]}{E(\eta^2) - \frac{2r}{n}E(\eta\delta) + (\frac{r}{n})^2E(\delta^2)} \tag{2.38}$$

It follows

$$\begin{aligned} B &= \frac{(\frac{r}{n})^2(\frac{1}{r}-\frac{1}{N})s_{xy} - \frac{r}{n}(\frac{1}{n}-\frac{1}{N})s_{xy}}{(\frac{1}{n}-\frac{1}{N})s_x^2 - \frac{2r}{n}(\frac{1}{n}-\frac{1}{N})s_x^2 + (\frac{r}{n})^2(\frac{1}{r}-\frac{1}{N})s_x^2} \\ &= \frac{(\frac{r}{n})^2(\frac{1}{r}-\frac{1}{N}) - \frac{r}{n}(\frac{1}{n}-\frac{1}{N})}{(\frac{1}{n}-\frac{1}{N}) - \frac{2r}{n}(\frac{1}{n}-\frac{1}{N}) + (\frac{r}{n})^2(\frac{1}{r}-\frac{1}{N})}(\frac{s_{xy}}{s_x^2}) \end{aligned} \tag{2.39}$$

(2.37) and (2.39) give the optimal coefficients A and B. However, noting $\bar{y}$ is the value that one just wants to estimate, it remains unknown during an imputation procedure. Consequently, since the optimal coefficient A is a function of $\bar{y}$, this method is not applicable theoretically. In practice, one can replace $\bar{y}$ by other estimation $\bar{y}_{imp}$ in (2.4), for instance, by $\bar{y}_{mean}$ in (2.16), or by $\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(x_i)$ found in Chapter 3.

2.2 Multiple Imputation Approach

Multiple imputation is a technique that tries to improve the single imputation method to resolve the problems of nonresponse data. In such an approach, one has a incomplete set of sample values Y and several complete sets of sample values $X^{(1)}, \cdots, X^{(m)}$ from the same population. As in Chapter 2.1, one has established the relations between the element in Y and the elements in each $X^{(k)}$, $k = 1, \cdots, m$. Now, one has interest to estimate a quantity Q, a function of the value set Y, in the survey. For each $k \in \{1, \cdots, m\}$, one obtains the corresponding estimator $\hat{Q}_{X^{(k)}}$ from X_k by a single imputation method. Then, one evaluates Q as a function, for instance, the mean, of these $\hat{Q}_{X^{(k)}}$.

In this chapter, we will give summary of the statistical theories given by Rubin [26] for the multiple imputation approach

to nonresponse in the census or survey.

2.2.1 Variables in the Multiple Imputation Approach

Here one defines four variables in the finite population of N individuals: X, covariates; Y, outcome variables; I, sampling indicators; and R, response indicators.

The covariates X describe characteristics of individuals that are fully observed or measured. X are written in the form:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1q} \\ X_{21} & X_{22} & \dots & X_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \dots & X_{Nq} \end{pmatrix} \tag{2.40}$$

where $(X_i) = \begin{pmatrix} X_{i1} & X_{i2} & \dots & X_{iq} \end{pmatrix}$ is a row vector that corresponds to the q components of covariate X, on the individual i, $i = 1, 2, \dots, N$

The outcome variables Y describe characteristics of individuals that are not fully observed or measured in the population. Y is written in the form:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1p} \\ Y_{21} & Y_{22} & \dots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & \dots & Y_{Np} \end{pmatrix} \tag{2.41}$$

where $(Y_i) = \begin{pmatrix} Y_{i1} & Y_{i2} & \dots & Y_{ip} \end{pmatrix}$ is a row vector corresponding to p components of characteristics of interest on the individual i.

The sampling indicators I describe which values are included or excluded in the survey. I can be written in the form:

$$I = \begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_N \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \dots & I_{1p} \\ I_{21} & I_{22} & \dots & I_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ I_{N1} & I_{N2} & \dots & I_{Np} \end{pmatrix} \tag{2.42}$$

where $(I_i) = \begin{pmatrix} I_{i1} & I_{i2} & \dots & I_{ip} \end{pmatrix}$ is a row vector. Each I_{ij} is defined by

$$I_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ recorded,} \\ 0 & \text{if } Y_{ij} \text{ not recorded.} \end{cases}$$

One assumes that the value of I_{ij} is known for all i and all j.

The response indicators R describe which values are respondent or nonrespondent in the survey. R can be written on the form:

$$R = \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_N \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} & \dots & R_{1p} \\ R_{21} & R_{22} & \dots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{N1} & R_{N2} & \dots & R_{Np} \end{pmatrix} \tag{2.43}$$

where $(R_i) = \begin{pmatrix} R_{i1} & R_{i2} & \dots & R_{ip} \end{pmatrix}$ is a row vector. Each R_{ij} is defined by

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ respondent} \\ 0 & \text{if } Y_{ij} \text{ nonrespondent.} \end{cases}$$

One assumes that the value of R_{ij} is known whenever $I_{ij} = 1$ and unknown whenever $I_{ij} = 0$.

In this thesis, we only consider the case $p = 1$, i.e., for every individual $i \in S$, the single response variable $Y_i = y_i \in \mathbb{R}$.

2.2.2 Analysis of Repeated Imputation

Let n be the sample size. Consider Q, be the quantity of interest in the survey. In order to estimate Q, one measures m complete sets of sample values $X^{(1)}, \cdots, X^{(m)}$. This yields m corresponding imputations $Y^{(1)}, \cdots, Y^{(m)}$ and thus m corresponding estimators $\hat{Q}^{(1)}, \cdots, \hat{Q}^{(m)}$ of Q. Assume that

$$Q - \hat{Q}^{(k)} \sim N(0, U^{(k)}) \qquad k = 1, \cdots, m \tag{2.44}$$

where $U^{(k)}$ is the variance of the estimator associated with $\hat{Q}^{(k)}$ and $N(0, U^{(k)})$ is the kth normal distribution with mean 0 and variance $U^{(k)}$. The estimates and associated variances for m sets of completed data can be combined as below.

The estimator for Q for multiple imputation is given by

$$\bar{Q} = \frac{1}{m}\sum_{k=1}^{m}\hat{Q}^{(k)} \tag{2.45}$$

The within-imputation variance is defined by

$$\bar{U} = \frac{1}{m}\sum_{k=1}^{m}U^{(k)} \tag{2.46}$$

and the between-imputation variance is defined by

$$B = \frac{1}{m-1}\sum_{k=1}^{m}(Q^{(k)} - \bar{Q})^t(Q^{(k)} - \bar{Q}) \tag{2.47}$$

The total variance of $Q - \bar{Q}$ is given by

$$T = \bar{U} + (1 + \frac{1}{m})B \tag{2.48}$$

Then, the statistic $T^{-1/2}(Q - \bar{Q})$ is approximately distributed as the Student t distribution on ν degree of freedom:

$$\frac{Q - \bar{Q}}{T^{1/2}} \sim t_\nu \tag{2.49}$$

with

$$\nu = (m-1)(1 + \frac{1}{r_m})^2 \tag{2.50}$$

where r_m, called *the relative increase in variance due to nonresponse*, is given by

$$r_m = \frac{(1 + m^{-1})B}{\bar{U}} \tag{2.51}$$

A $100(1-\alpha)\%$ confidence interval of estimate Q is then found as

$$\bar{Q} \pm t_\nu(\alpha/2)T^{1/2} \tag{2.52}$$

2.2.3 Multiple Imputation Efficiency

One notes, by (2.50), that the degrees of freedom ν depends on the repeated number m and the ratio r_m. When m increases or r_m decreases, the degrees of freedom ν increases and thus, the statistic $T^{-1/2}(Q-\bar{Q})$ tends to be distributed as a normal distribution. When there are no missing data about Q, by (2.47), one sees that $B=0$, hence, $r_m=0$, the distribution in (2.49) will be normal as previewed.

Another estimate of the fraction of missing data about Q due to nonresponse, derived by Rubin ([26] p.77), is the rate of missing information:

$$\gamma_m = \frac{r_m + 2/(\nu+3)}{r_m+1} \tag{2.53}$$

Rubin showed that the efficiency, in units of variance, of finite-m imputation estimator relative to the infinite-m imputation

estimator is approximately given by ([26] p.114)

$$RE = (1 + \frac{\gamma_0}{m})^{-1} \tag{2.54}$$

where γ_0 is the population fraction of missing information.

By calculating ([26] p.114) the efficiency RE in terms of different values of m and γ_0, Rubin claimed that $m = 2$ to 10 imputations may be proper.

2.2.4 Evaluation of Multiple Imputation Method

One should note that no technique can be perfect or unimportant for the nonresponse problem. The multiple imputation technique has its own advantages and inconveniences.

Conditions of Advantage

Multiple imputation retains the ability of single imputation to use the complete data method of analysis. And it enhances the ability of single imputation to incorporate the data collector's knowledge because the data collectors are allowed to use their knowledge to reflect uncertainty about which values will be imputed. Multiple Imputation always produces estimates which are more representative of the population than other popular methods of handling mising data. In addition to the shared

advantage with single imputation, there are some distinct advantages.

- Multiple Imputation increases the efficiency of estimation.
- Valid inference which reflect the additional variability due to missing values can be simply obtained by straightforward combining completed data inferences under a model.

Conditions of Inconvenience

Obviously there are some disadvantages of multiple imputation relative to single imputation.

- More work is required to produce sets of completed data.
- More storage space is needed to store these multiple imputed data sets.
- Extra work is required to analyze these multiple imputed data sets.

Chapter 3

Nonparametric Imputation Approaches

As in Section 2.1 and Section 2.2, one has two correlated sets of independent and identically distributed random vectors: the complete set X and the incomplete set Y from a sample survey $S = \{1, \cdots, n\} \subseteq \Omega$, the set of the population. The object here is to establish a regression curve by these two sets which provides a reasonable estimation to the representation of Y. We call such regression technique as a *smoothing*.

Let $Z_S = \{(X_i, Y_i) : X_i \in X_S, Y_i \in Y_S\}$, where $X_S = \{X_1, \cdots, X_n\}$ and $Y_S = \{Y_1, \cdots, Y_n\}$ are the set of the covariates and the set of outcome variables respectively from the sample S defined in Section 2.1 and Section 2.2.

In this thesis, we only consider the case where for every individual $i \in S$, the single response variable $Y_i = y_i \in \mathbb{R}$ and the

predictor variable $X_i \in \mathbb{R}^q$.

Let $R_S = \{R_i : i \in S\}$ be the response indicator defined by

$$R_i = \begin{cases} 1 & \text{if } Y_i \text{ respondent,} \\ 0 & \text{if } Y_i \text{ nonrespondent.} \end{cases} \tag{3.1}$$

The nonparametric imputation assumes an appropriate regression relationship

$$Y = m(X) + \epsilon \tag{3.2}$$

so that ,

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \cdots, n \tag{3.3}$$

To obtain the regression curve $m(x)$, one defines an estimator of $m(x)$ by

$$\hat{m}(x) = \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} W_i(x) R_i Y_i \tag{3.4}$$

where $W_i(x)$, depending on the vectors $X_1, \cdots, X_n$, is the weight of Y_i to the individual $i \in S$. We call such regression estimator $\hat{m}(x)$ as a *smoother* and the corresponding outcome $\{\hat{Y}_i = \hat{m}(X_i) : i \in S\}$ as the *smooth value*. Thus, the smoothing of $Z_S = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ becomes a procedure of how to find these weights $W_i(x)$ for every individual $i \in S$.

In this chapter, we will summarize some smoothing techniques

given by Härdle [12].

3.1 Kernel Smoothing

In this section, the predictor variables $X_i \in \mathbb{R}$, i.e., x is one-dimensional real scalar.

In order to find the weight sequence $\{W_1(x), \cdots, W_n(x)\}$, one introduces a continuous, bounded and symmetric real function K called the *Kernal function*, such that

$$\int_{\mathbb{R}} K(t)dt = 1 \tag{3.5}$$

The idea of the kernel smoothing is motivated from the properties of a probability density function $f(x)$. The equation (3.5) is, in fact, a feature that a probability density function must satisfy. Let $F(x) = P(X \leq x)$ be the cumulative distribution function of X.

$$f(x) = \frac{dF}{dx} = \lim_{h \to 0} \frac{P(x - h < x < x + h)}{2h} \tag{3.6}$$

This gives an idea to estimate $f(x)$, from the n observations

$X_1, \cdots, X_n$, by ([8] p.324)

$$\begin{aligned}\hat{f}_n(x) &\stackrel{def}{=} \frac{1}{2h}\Big[\frac{\text{number of observations falling in}(x-h, x+h)}{n}\Big] \\ &= \frac{1}{hn}\sum_{i=1}^{n} w\Big(\frac{x - X_i}{h}\Big)\end{aligned} \tag{3.7}$$

where w the weight function defined by

$$w(x) = \frac{1}{2} I_{\{|x|\leq 1\}} \tag{3.8}$$

Note that if $K(x)$ is the density function of the random variable X, then

$$K_h(x) \stackrel{def}{=} \frac{1}{h} K(\frac{x}{h}) \tag{3.9}$$

is the density function of the random variable hX. With the kernel equation (3.5), Rosenblatt and Parzen ([8] p.331) defined a *kernel* density estimator for the density function of X by

$$\hat{f}_h(x) = \frac{1}{hn}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) \tag{3.10}$$

Since K is continuous, bounded, symmetric about the origin on real line, and integrated to unity by (3.5), then for a small h, the regression relation m at some point x_0 in (3.2) can be written

approximately as

$$\begin{aligned} m(x_0) &\approx \int_{-\infty}^{\infty} m(t) K_h(t - x_0) dt \\ &= \mathbb{E}\Big[\frac{m(X) K_h(X - x_0)}{f(X)}\Big] \\ &= \mathbb{E}\Big[\frac{Y K_h(X - x_0)}{f(X)}\Big] \end{aligned} \tag{3.11}$$

From the equation (3.11), a kernel estimator can be obtained as:

$$\hat{m}_h(x) = \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} \frac{[K_h(X_i - x) R_i Y_i]}{f(X_i)} \tag{3.12}$$

Usually the density function of $f(x)$ is unknown. Note that the kernel density (3.10) is enlightened by the naive density estimator (3.7). For very small h, these $f(X_i)$ in (3.12) can be replaced by $\hat{f}_h(x)$. Thus, the estimator (3.12) becomes

$$\hat{m}_h(x) = \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} \frac{[K_h(x - X_i) R_i Y_i]}{\hat{f}_h(x)} \tag{3.13}$$

Consequently, Nadaraya and Watson ([12] p.25) proposed the weight sequence of the kernel smoother as

$$\begin{aligned} W_i(x) &= \frac{K_h(x - X_i)}{\hat{f}_h(x)} = \frac{K_h(x - X_i)}{\frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j)} \\ &= \frac{K(\frac{x - X_i}{h})}{\frac{1}{n}\sum_{j=1}^{n} K(\frac{x - X_j}{h})} \end{aligned} \tag{3.14}$$

where the size of the weight h is called the *bandwidth*. This gives the *Nadaraya-Watson estimator* by (3.4):

$$\begin{aligned}\hat{m}_h(x) &= \frac{1}{\sum_{k=1}^{n} R_k} \sum_{i=1}^{n} \Big[\frac{K_h(x - X_i) R_i Y_i}{\frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j)} \Big] \\ &= \frac{\sum_{i=1}^{n} [K(\frac{x-X_i}{h}) R_i Y_i]}{\frac{\sum_{k=1}^{n} R_k}{n} \sum_{j=1}^{n} K(\frac{x-X_j}{h})}\end{aligned} \tag{3.15}$$

We note that if $R_i = 1$ for all $i \in S$, i.e., all Y_i are respondent, (3.15) becomes

$$\begin{aligned}\hat{m}_h(x) &= \frac{1}{n} \sum_{i=1}^{n} \Big[\frac{K_h(x - X_i) Y_i}{\frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j)} \Big] \\ &= \frac{\sum_{i=1}^{n} [K(\frac{x-X_i}{h}) Y_i]}{\sum_{j=1}^{n} K(\frac{x-X_j}{h})}\end{aligned} \tag{3.16}$$

The above equation implies a reasonable estimator in the case of missing data, introduced by Cheng and Wei [5],

$$\begin{aligned}\hat{m}_h(x) &= \sum_{i=1}^{n} \Big[\frac{K_h(x - X_i) R_i Y_i}{\sum_{j=1}^{n} K_h(x - X_j) R_j} \Big] \\ &= \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h}) R_i Y_i}{\sum_{j=1}^{n} K(\frac{x-X_j}{h}) R_j}\end{aligned} \tag{3.17}$$

That is obtained by taking account of only these variables X_i for $R_i = 1$, i.e., the kernel density estimator (3.10) for the density function of X becomes

$$\hat{f}_h(x) = \frac{1}{h \sum_{i=1}^{n} R_i} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) R_i \tag{3.18}$$

and thus the weight sequence of the kernel smoother (3.14) becomes

$$\begin{aligned} W_i(x) &= \frac{K_h(x - X_i)}{\hat{f}_h(x)} = \frac{K_h(x - X_i)}{\frac{1}{\sum_{k=1}^{n} R_k} \sum_{j=1}^{n} K_h(x - X_j) R_j} \\ &= \frac{K(\frac{x - X_i}{h})}{\frac{1}{\sum_{k=1}^{n} R_k} \sum_{j=1}^{n} K(\frac{x - X_j}{h}) R_j} \end{aligned} \tag{3.19}$$

Substituting (3.19) into (3.4), we obtain (3.17).

3.2 Selection of Kernel Function

In this section, we suppose that $R_i = 1$ for all $i \in S$, i.e., all Y_i are respondent.

Let

- K be the kernel function as defined in Section 3.1.
- X be a one-dimensional predictor variable.
- m be the regression model defined by (3.2).
- f be the density function of the random variable X.
- $\sigma^2(x)$ be the variance of the random variable X at the point x.

Then one has the following theorem ([12] p.29).

Theorem 3.1 *Assume:*

1. $\int_{\mathbb{R}}|K(t)|dt < \infty$*;*

2. $\lim_{|t|\to\infty} tK(t) = 0$*;*

3. $\mathbb{E}[Y^2] < \infty$*;*

4. for $n \to \infty$*, then,* $h \to 0$ *and* $nh \to \infty$*.*

Then, for each $x \in \mathbb{R}$ *with* $f(x) > 0$ *such that* m*,* f *and* σ^2 *are continue at the point* x*, one has*

$$\frac{1}{n}\sum_{i=1}^{n} W_i(x)Y_i \xrightarrow{P} m(x).$$

The theorem 3.1 ensures that the kernel smoother $\hat{m}_h(x)$ converges in probability to response curve $m(x)$. Based on the Theorem 3.1, one can propose several kernel functions K to estimate the response variable Y.

Example 3.2.1 *Polynomial kernels*

$$\begin{aligned}
(1)\quad & K(t) = \frac{3}{4}(1-t^2)I(|t| \le 1) \\
(2)\quad & K(t) = \frac{15}{32}(3-10t^2+7t^4)I(|t| \le 1) \\
(3)\quad & K(t) = \frac{15}{4}(-t+t^3)I(|t| \le 1) \\
(4)\quad & K(t) = \frac{105}{32}(-5t+14t^3-9t^5)I(|t| \le 1) \\
(5)\quad & K(t) = \frac{105}{16}(-1+6t^2-5t^4)I(|t| \le 1) \\
(6)\quad & K(t) = \frac{315}{64}(-5+63t^2-135t^4+77t^6)I(|t| \le 1)
\end{aligned} \tag{3.20}$$

In [12], Härdle listed some polynomial kernels ([12] p.135. (1) is also called as the Epanechnikov kernel). All of these kernel functions in Example 3.2.1 have support [-1, 1] and were derived by Gasser, Müller and Mammitzsch [1985] from some optimalily consideration.

Example 3.2.2 *The Gaussian kernel*

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{3.21}$$

In fact, is the density of a standard normal distribution.

Example 3.2.3 *The gamma kernel*

$$K(t) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} e^{-\frac{t}{\beta}} t^{\alpha-1} \qquad t \geq 0 \tag{3.22}$$

is, in fact, the density of the gamma distribution. Note that in Example 3.2.3, the gamma kernel K is not symmetric to the origin 0. Ignoring the condition of symmetry, the density function $f(x)$ is a kernel function of the continuous random variable X, where α and β have to be chosen appropriately.

Note that the function K in Example 3.2.3 does not satisfy the condition of symmetry for a kernel. Thus, for using (3.22) as a kernel function, it should have some special consideration.

3.3 Selection of Bandwidth

In this section, we suppose that $R_i = 1$ for all $i \in S$, i.e., all Y_i are respondent.

Let

$$d_M(x,h) = \mathbb{E}[\hat{m}_h(x) - m(x)]^2$$

be the mean squared error of the kernel smoothing model at a point x of the random variable X. For the regression method (3.2), assuming, without loss of generality, that X_i is taken from the interval $[-1,1]$, Gasser and Müller ([12] p.29) in 1984 showed the following theorem .

Theorem 3.2 *Let*

$$c_K = \int_{\mathbb{R}} K^2(t)dt < \infty$$
$$d_K = \int_{\mathbb{R}} t^2 K(t)dt < \infty$$

Take the kernel weight sequence $\{W_{hi}\}$*, proposed by Gasser and Müller([12] p.28), as*

$$W_{hi} = n \int_{S_{i-1}}^{S_i} K_h(x-t)dt \tag{3.23}$$

where $X_{i-1} \leq S_{i-1} \leq X_i$ *is chosen from the ordered set* $\{X_0 < X_1 <, \cdots, < X_n\}$ *with* $X_0 = -1$.

Assume:

1. K *has support [-1, 1] with* $K(-1) = K(1) = 0$*;*

2. $m \in C^2$*;*

3. $max_i|X_i - X_{i-1}| = O(n^{-1})$*;*

4. $var(\epsilon_i) = \sigma^2, i = 1, \cdots, n$*;*

5. for $n \to \infty$*, then,* $h \to 0$ *and* $nh \to \infty$*. Then, for each* $x \in \mathbb{R}$ *with* $f(x) > 0$ *such that* m*,* f *and* σ^2 *are continue at the point* x*, one has*

$$d_M(x,h) \approx (nh)^{-1}\sigma^2 c_K + h^4 d_K^2 (m''(x))^2/4.$$

The optimal bandwidth h_{opt} can be found by setting

$$\frac{\partial d_M(x,h)}{\partial h} = 0$$

3.4 Kernel Smoothing for Non-negative Stationary Ergodic Processes

In this section we suppose that $R_i = 1$ for all $i \in S$, i.e., all Y_i are respondent.

In Section 3.1, we have shown the idea how to get the Nadaraya-Watson (NW) estimator (3.15):

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)} \tag{3.24}$$

where K_h, represents the density function of the random variable hX, is given by (3.9). In this section, we summarize the kernel smoothing, proposed by Chaubey, Laïb and Sen in [3],

to give a rigor and solid reason to the generalized NW estimator from (3.24) for non-negative data sampled from a stationary, ergodic process. This says that for every individual $i \in S$, the single response variable $Y_i \in \mathbb{R}^+$ and the predictor variable $X_i \in \mathbb{R}^{+q}$, and furthermore, every $Z_i = (X_i, Y_i) \in Z_S = \{(X_i, Y_i) : X_i \in X_S, Y_i \in Y_S\}$ is sampled from a stationary, ergodic process.

Define the conditional mean function

$$m(x) = \mathbb{E}(\phi(Y_1)|X_1 = x) \tag{3.25}$$

where $\phi : \mathbb{R}^+ \to \mathbb{R}$, is a *Borel* function such that $\mathbb{E}(|\phi(Y_1)|) < \infty$, in addition, $m(x) = \mathbb{E}(\phi(Y_1)|X_1 = x) < \infty$ for any $x \in X_S$. The goal in this section is to construct an estimator for the mean function m.

3.4.1 1-Dimensional Case

Here we consider $X_i \in \mathbb{R}^+$ for every $i \in S$, i.e., $q = 1$.

In order to construct the estimator of m in (3.25), Chaubey, Laïb and Sen presented the following theorem ([3] p.975), originally shown in [11], Chapter VII ([11] p.219).

Theorem 3.3 *For $n = 1, 2, \cdots$, consider a family of distributions f_{x,h_n}, where h_n denotes the bandwidth for each sample S_n, with mean $\mu_n(x)$ and variance $\sigma_n^2(x)$. Let u be a bounded and continuous function on $\mathbb{R}$. If $\mu_n(x) \to x$ and $\sigma_n^2(x) \to 0$ for each $x \in \mathbb{R}$, then*

$$\int_{-\infty}^{\infty} u(t) f_{x,h_n}(t) dt \to u(x) \quad as \quad n \to \infty \tag{3.26}$$

The convergence is uniform in every finite interval in which $\sigma_n^2(x) \to 0$ uniformly and u is uniformly continuous.

Let $u(t) = m(t)f(t)$, where f, bounded and continuous on $\mathbb{R}^+$, is the common density function of $X_i \in X_S$. Then, (3.26) becomes

$$\mathbb{E}_f(\phi(Y_1) f_{x,h_n}(X_1)) = \int_{-\infty}^{\infty} m(t) f(t) f_{x,h_n}(t) dt \to m(x) f(x) \quad \text{as} \quad n \to \infty \tag{3.27}$$

(3.27) motivates an estimator of m for each sample S_n given by :

$$\hat{m}_n(x) = \frac{\hat{u}_n(x)}{\hat{f}_n(x)}$$

for $x \in \mathbb{R}^+$, where $\hat{u}_n$ is the estimator of $\mathbb{E}_f(\phi(Y_1) f_{x,h_n}(X_1))$:

$$\hat{u}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i) f_{x,h_n}(X_i) \tag{3.28}$$

and $\hat{f}_n$ is the estimator of f:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} f_{x,h_n}(X_i) \tag{3.29}$$

i.e., the estimator of m for each sample S_n is constructed by

$$\hat{m}_n(x) = \frac{\sum_{i=1}^{n} \phi(Y_i) f_{x,h_n}(X_i)}{\sum_{i=1}^{n} f_{x,h_n}(X_i)} \tag{3.30}$$

We note that the Nadaraya-Watson estimator $\hat{m}_h$ in (3.24) is a special case of the estimator $\hat{m}_n$ given in (3.30) by taking the bandwidth $h_n = h$ and

$$f_{x,h_n}(t) = K_h(x-t) = \frac{1}{h}K\Big(\frac{x-t}{h}\Big)$$

where K is the kernel described in Section 3.1.

Chaubey, Laïb and Sen indicated ([3] p.975) that the estimator in (3.30) may not be defined at $x = 0$ except in cases where $\hat{m}_n(0) = \lim_{x\to 0^+} \hat{m}_n(x)$ exists. For this situation, Chaubey, Laïb and Sen modified the estimator in (3.30) to construct a *perturbed* version of estimator of m ([3] p.976):

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n \phi(Y_i) f_{x+\epsilon_n,h_n}(X_i)}{\sum_{i=1}^n f_{x+\epsilon_n,h_n}(X_i)} \tag{3.31}$$

where $\epsilon_n \in \mathbb{R}^+$ tends to 0 as $n \to \infty$ in an appropriate rate.

3.4.2 q-Dimensional Case

Here we consider $X_i \in \mathbb{R}^{+q}$ with $q > 1$ for every $i \in S$. In this case, the estimator of m is still given by (3.31), i.e.,

$$\hat{m}_n(x) = \frac{\hat{u}_n(x+\epsilon_n)}{\hat{f}_n(x+\epsilon_n)} = \frac{\sum_{i=1}^n \phi(Y_i) f_{x+\epsilon_n,h_n}(X_i)}{\sum_{i=1}^n f_{x+\epsilon_n,h_n}(X_i)} \tag{3.32}$$

In this multi dimensional case, usually these estimators of distributions f_{x,h_n} can be calculated by

$$f_{x,h_n}(t) = \prod_{k=1}^{q} f_{x_k,h_n}(t)$$

where $x = (x_1, \cdots, x_q) \in \mathbb{R}^{+q}$, and f_{x_k,h_n} denotes the distribution described in the theorem 3.3 for the element $x_k, k \in \{1, \cdots, q\}$.

In [3], Chaubey, Laïb and Sen studied the specific kernel estimator given by a Gamma distribution. They showed, under certain conditions, the results of asymptotic normality for the estimator and evaluated the mean squared errors (MSE).

3.5 Nearest Neighbor Estimates

The nearest neighbor imputation is to use a *k-nearest neighbor (k-NN)* sequence $\{(X_i, Y_i)\}_{i=1}^n$, introduced by Loftsgaarden and Quesenberry [1965], to estimate the smoother $m(x)$ at the point x in the regression relationship (3.2). In order to calculate the smoother $m(x)$ by the *k-NN* imputation, the k observations X_i closest to x are chosen for the index set

$J_x = \{i : X_i$ is one of the k nearest observations with $R_i = 1$ to $x\}$

and the *k-NN* weight sequence is constructed by ([12] p.42)

$$W_{ki}(x) = \begin{cases} 1/k & \text{if } i \in J_x, \\ 0 & \text{otherwise .} \end{cases} \tag{3.33}$$

Then, the *k-NN* smoother is found by (3.4):

$$\hat{m}_k(x) = \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} W_{ki}(x) R_i Y_i \tag{3.34}$$

Let X be a one-dimensional predictor variable. Suppose that $R_i = 1$ for all $i \in S$, i.e., all Y_i are respondent. For the *k-NN* estimate, one has the following theorem ([12] p.43).

Theorem 3.4 *Let $k \to \infty$, $k/n \to \infty$, $n \to \infty$. The bias and variance of the k-NN estimate $\hat{m}_k$ with weights as in (3.34) are given by*

$$\mathbb{E}[\hat{m}_k(x) - m(x)] \approx \frac{1}{24 f(x)^3}[(m'' f + 2m' f')(x)](k/n)^2$$

$$var\{\hat{m}_k(x)\} \approx \frac{\sigma^2(x)}{k}$$

where m is the regression model defined by (3.2), f is the density function of the random variable X, and $\sigma^2(x)$ is the variance of the random variable X at the point x.

3.6 Horvitz-Thompson Estimate

Let $Z_i = (X_i, Y_i) \in Z_S = \{(X_i, Y_i) : X_i \in X_S, Y_i \in Y_S\}$. The population total is defined by

$$\tau = \sum_{i=1}^{N} Y_i \tag{3.35}$$

In 1952, Horvitz and Thompson proposed an inverse weighting estimator of the population total τ ([6] p.259):

$$\hat{\tau} = \sum_{i=1}^{n} \frac{R_i Y_i}{\pi_i} \tag{3.36}$$

where π_i is the probability that the ith unit Y_i is in the sample:

$$\pi_i = \mathbb{P}(R_i = 1 | X_i \in X_S) \tag{3.37}$$

Theorem 3.5 *If* $\pi_i > 0$, $i = 1, \cdots, n$,

$$\hat{\tau} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}$$

is an unbiased estimator of τ, *with variance*

$$var(\hat{\tau}) = \sum_{i=1}^{n} \frac{1 - \pi_i}{\pi_i} Y_i^2 + 2 \sum_{i}^{n} \sum_{j>i}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j$$

where π_{ij} is the probability that the ith and jth units Y_i, Y_j are both in the sample:

$$\pi_{ij} = \mathbb{P}(R_i = 1, R_j = 1 | X_i, X_j \in X_S) \tag{3.38}$$

3.7 Nonparametric Regression Imputation Method to Estimate the Population Mean

The nonparametric regression weighting approach is a common way to impute the missing values for the analysis of nonparamet-

ric imputation in sampling techniques. The regression approach includes :

- Kernel regression imputation
- Nearest neighbor imputation

Here we summarize some kernel smoothing estimators, proposed by Ning and Cheng in [20].

3.7.1 Kernel Regression Weighting Method

By (3.17), Cheng and Wei [5] introduced an estimator to the population mean for Y_i:

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\sum_{k=1}^{n}\hat{m}_h(X_k) \\ &= \frac{1}{n}\sum_{k=1}^{n}\frac{\sum_{i=1}^{n}K_h(X_k - X_i)R_iY_i}{\sum_{j=1}^{n}K_h(X_k - X_j)R_j} \\ &= \frac{1}{n}\sum_{k=1}^{n}\frac{\sum_{i=1}^{n}K(\frac{X_k - X_i}{h})R_iY_i}{\sum_{j=1}^{n}K(\frac{X_k - X_j}{h})R_j}\end{aligned} \tag{3.39}$$

A reasonable estimator to the population mean for Y_i is

$$\hat{\mu}_{KR} = \frac{1}{n}\sum_{i=1}^{n}[R_iY_i + (1 - R_i)\hat{m}_h(X_i)] \tag{3.40}$$

where $\hat{m}_h(x)$ is given by (3.17). In (3.40), the nonrespondent Y_i are replaced by $\hat{m}_h(X_i)$.

3.7.2 Nearest Neighbor Regression Weighting Method

The *nearest neighbor (NN) regression weights* is an another basic approach to nonparametric imputation. For a finite positive interger k,a *k-NN* imputation estimator is defined as

$$\hat{\mu}_{NN} = \frac{1}{n}\sum_{i=1}^{n}[R_i Y_i + (1 - R_i)\hat{m}_k(X_i)] \tag{3.41}$$

where $\hat{m}_k$ is the *k-NN* smoother given by (3.34). Similar to (3.40), in (3.41) the nonrespondent Y_i are replaced by $\hat{m}_k(X_i)$.

3.7.3 Horvitz-Thompson(HT) Inverse Weighting Method

According to (3.36), the HT estimator of the population mean μ is defined by

$$\hat{\mu}_{HT} = \frac{\hat{\tau}}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi_i} \tag{3.42}$$

By similar reasoning to construct the estimator of population total (3.36) ([6] p.259), one can replace n in (3.42) by a ratio estimator:

$$\hat{n} = \sum_{i=1}^{n}\frac{R_i}{\pi_i} \tag{3.43}$$

Then the HT estimator of the population mean (3.42) becomes

$$\hat{\mu}_{HT} = \frac{1}{\sum_{i=1}^{n}\frac{R_i}{\pi_i}}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi_i} \tag{3.44}$$

According to (3.37), π_i is a conditional probability. By the law of conditional probability and the kernel density estimator $\hat{f}_h(x)$ for the density function of X (3.10), one obtains an estimator of π_i:

$$\hat{\pi}_i = \frac{\hat{f}_h(R_i = 1, X_i)}{\hat{f}_h(X_i)} = \frac{\sum_{j=1}^{n} R_i K_h(X_i - X_j)}{\sum_{j=1}^{n} K_h(X_i - X_j)} \tag{3.45}$$

where K_h is the kernel given by (3.9).

3.8 Multiple Imputation

In this sections, we consider the case where the predictor variable X is $p-$dimensional, i.e., $X_i = (X_{i1}, \cdots, X_{ip})$ for every individual $i \in S = \{1, \cdots, n\}$. In such multi dimensional case, the kernel function can be defined as

$$K(t_1, \cdots, t_p) = \prod_{k=1}^{p} K(t_k) \tag{3.46}$$

where $K(t_k)$ is the one-dimensional kernel function defined in Section 3.1.

Then the kernel weights can be found by

$$W_i(x) = \frac{\prod_{k=1}^{p} K_{h_k}(x_k - X_{ik})}{\hat{f}_h(x)} \tag{3.47}$$

where K_h is defined by (3.9), and $\hat{f}_h$ is the Rosenblatt-Parzen density estimator defined by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}(\prod_{k=1}^{p} K_{h_k}(x_k - X_{ik})) \tag{3.48}$$

By (3.4), one obtains an estimator for Y:

$$\begin{aligned}\hat{m}_h(x) &= \frac{\sum_{i=1}^{n}\prod_{k=1}^{p} K_{h_k}(x_k - X_{ik})R_iY_i}{\frac{\sum_{i=1}^{n} R_i}{n}\sum_{i=1}^{n}(\prod_{k=1}^{p} K_{h_k}(x_k - X_{ik}))} \\ &= \frac{\sum_{i=1}^{n}\prod_{k=1}^{p} K(\frac{x_k - X_{ik}}{h_k})R_iY_i}{\frac{\sum_{i=1}^{n} R_i}{n}\sum_{i=1}^{n}(\prod_{k=1}^{p} K(\frac{x_k - X_{ik}}{h_k}))}\end{aligned} \tag{3.49}$$

Similarly with (3.17), one can construct another reasonable estimator for Y:

$$\begin{aligned}\hat{m}_h(x) &= \frac{\sum_{i=1}^{n}\prod_{k=1}^{p} K_{h_k}(x_k - X_{ik})R_iY_i}{\sum_{i=1}^{n}(\prod_{k=1}^{p} K_{h_k}(x_k - X_{ik}))R_i} \\ &= \frac{\sum_{i=1}^{n}\prod_{k=1}^{p} K(\frac{x_k - X_{ik}}{h_k})R_iY_i}{\sum_{i=1}^{n}(\prod_{k=1}^{p} K(\frac{x_k - X_{ik}}{h_k}))R_i}\end{aligned} \tag{3.50}$$

Chapter 4

Application

In this chapter, we apply these approaches introduced in Chapter 2 and Chapter 3 to a real example: a coast to coast chain of stores and their supplies in Canada (The data, given by Professor Wei Sun of Department of Mathematics and Statistics of Concordia University, are quoted from the author's report of the B Sc.Honors project in 2011).

The data table in Appendix A consists of the work hours and sales in the first week and fifth week of a year by 287 divisions of the stores in the chain. We believe that the work hours of different weeks, the work hours and sales in same week, the sales of different weeks are strongly correlated. The data in Appendix A are completed. We delete almost 100 data of sales in the fifth week. Then, we use the methods introduced in Chapter 2 and Chapter 3 to estimate these missing data and verify the applicability of those methods.

The programs we made here (The programs were made by Matlab). can be used for the similar general cases. For instance, the heights and weights of a groups of boys at age 15 in a certain region were measured, the approaches applied here can be used to estimate the missing data of heights and weights for the same groups at age 20 and predict the average height and weight for the 20-year-order boys in the same region.

4.1 Single Imputation

In this section, $X_S = \{X_1, \cdots, X_n\}$ is the sample of work hours in week 1, $Y_S = \{Y_1, \cdots, Y_n\}$ is the sample of work hours in week 5 for these divisions of the stores, where $n = 287$, and $R = \{R_1, \cdots, R_n\}$ is the response indicator set defined in (2.1).

4.1.1 Ratio Method of Imputation

The estimator of Y_k is given by (2.20):

$$\hat{Y}_k = \hat{m}(X_k) = \begin{cases} Y_k & \text{if } R_k = 1 \\ \frac{\bar{Y}_r}{\bar{X}_r} X_k & \text{if } R_k = 0 \end{cases} \tag{4.1}$$

where

$$\bar{X}_r = \frac{1}{\sum_{j=1}^{n} R_j} \sum_{j=1}^{n} R_j X_j$$

$$\bar{Y}_r = \frac{1}{\sum_{j=1}^{n} R_j} \sum_{j=1}^{n} R_j Y_j$$

The estimator is illustrated in Fig. 4.1.

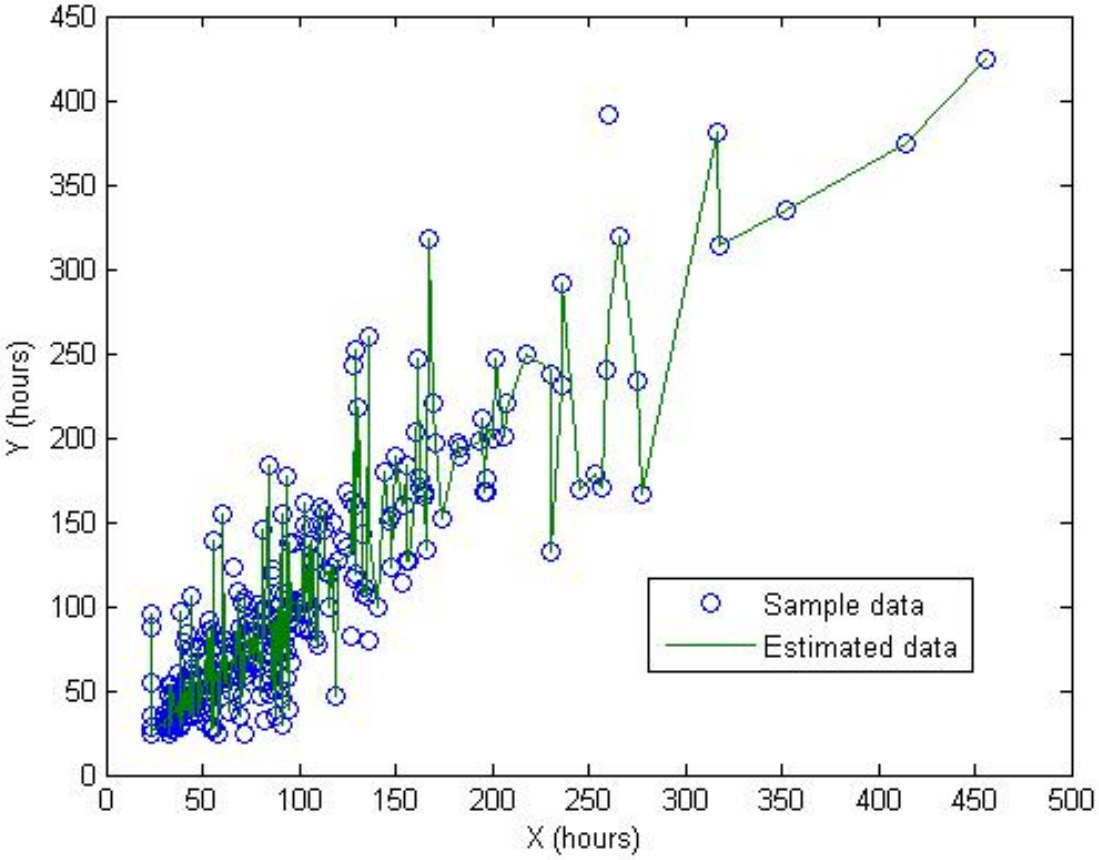

Figure 4.1: Ratio estimator

The estimation of mean $\bar{Y}$ is given by (2.21):

$$\begin{aligned} \bar{Y}_{ratio} &= \frac{1}{n}\sum_{k=1}^{n}\hat{Y}_k \\ &= \bar{Y}_r(\frac{\bar{X}_n}{\bar{X}_r}) \end{aligned} \tag{4.2}$$

where $\bar{X}_n$ is the average value of the sample X_S:

$$\bar{X}_n = \bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

We obtain:

$$\bar{Y}_{ratio} \approx 100.8651 \text{ hours}$$

and

$$\begin{aligned} \text{Error } &= |\bar{Y}_{ratio} - \bar{Y}| \approx 0.9783 \text{ hours} \\ \text{Relative error } &= |\frac{\bar{Y}_{ratio} - \bar{Y}}{\bar{Y}}| \approx 0.0096 \end{aligned}$$

where

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \approx 101.8434 \text{ hours} \tag{4.3}$$

is the average of the work hours in the fifth week from the sample given by the data table in Appendix A.

4.1.2 Regression Method of Imputation

The estimator of Y_k is given by (2.26):

$$\hat{Y}_k = \begin{cases} Y_k & \text{if } R_k = 1 \\ \bar{Y}_r - \frac{s_{XY}}{s_X^2}\bar{X}_r + \frac{s_{XY}}{s_X^2}X_k & \text{if } R_k = 0 \end{cases} \tag{4.4}$$

where s_X, s_Y, s_{XY} are given by (2.5):

$$s_{XY} = \frac{1}{r-1}\sum_{k=1}^{n}(R_k X_k - \bar{X}_r)(R_k Y_k - \bar{Y}_r)$$

$$s_X^2 = \frac{1}{r-1}\sum_{k=1}^{n}(R_k X_k - \bar{X}_r)^2$$

$$s_Y^2 = \frac{1}{r-1}\sum_{k=1}^{n}(R_k Y_k - \bar{Y}_r)^2$$

with

$$r = \sum_{j=1}^{n} R_j$$

The estimator is illustrated by Fig. 4.2.

The estimation of mean $\bar{Y}$ is given by (2.27):

$$\begin{aligned} \bar{Y}_{regression} &= \frac{1}{n}\sum_{k=1}^{n}\hat{Y}_k \\ &= \bar{Y}_r - \frac{s_{XY}}{s_X^2}\bar{X}_r + \frac{s_{XY}}{s_X^2}\bar{X}_n \end{aligned} \tag{4.5}$$

We obtain:

$$\bar{Y}_{regression} \approx 101.3384 \text{ hours}$$

and

$$\text{Error } = |\bar{Y}_{regression} - \bar{Y}| \approx 0.5050 \text{ hours}$$

$$\text{Relative error } = |\frac{\bar{Y}_{regression} - \bar{Y}}{\bar{Y}}| \approx 0.0050$$

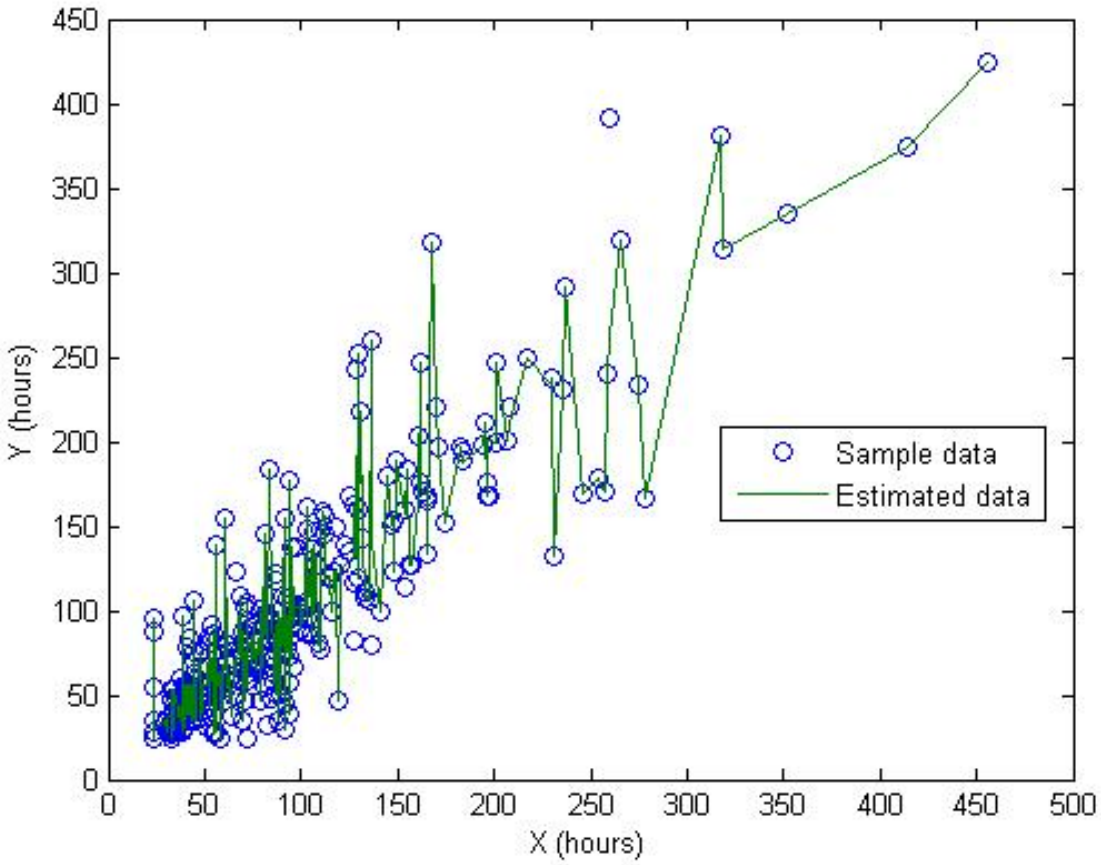

Figure 4.2: Regression estimator

4.1.3 Optimal Method of Imputation

The estimator of Y_k is given by (2.32):

$$\hat{Y}_k = \begin{cases} Y_k & \text{if } R_k = 1 \\ A + BX_k & \text{if } R_k = 0 \end{cases} \tag{4.6}$$

where

$$B = \frac{s_{XY}}{s_X^2}$$

$$A = \hat{\bar{Y}} - B\bar{X} = \bar{Y}_r - B\bar{X}$$

The estimator is illustrated by Fig. 4.3.

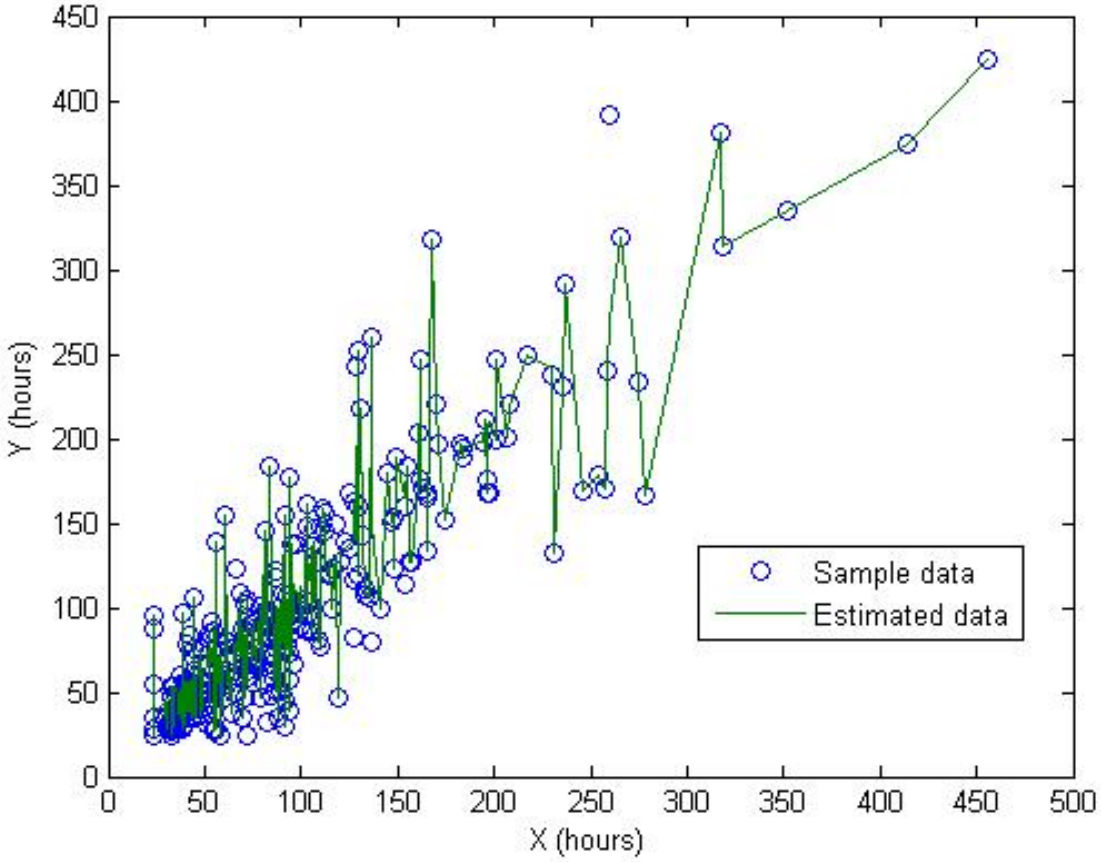

Figure 4.3: Optimal estimator

The estimation of mean $\bar{Y}$ is given by (2.32):

$$\begin{aligned} \bar{Y}_{optimal} &= \frac{1}{n}\sum_{k=1}^{n}\hat{Y}_k \\ &= \frac{r}{n}\bar{Y}_r + (1-\frac{r}{n})A + B(\bar{X}_n - \frac{r}{n}\bar{X}_r) \end{aligned} \tag{4.7}$$

We obtain:

$$\bar{Y}_{optimal} \approx 104.2080 \text{ hours}$$

and

$$\begin{aligned} \text{Error } &= |\bar{Y}_{optimal} - \bar{Y}| \approx 2.3646 \text{ hours} \\ \text{Relative error } &= |\frac{\bar{Y}_{optimal} - \bar{Y}}{\bar{Y}}| \approx 0.0232 \end{aligned}$$

4.1.4 Kernel Smoothing Method

In this section, we apply those approaches introduced in Section 3.2 to estimate the work hours in the fifth week. The estimator of Y_k is given by (3.17):

$$\hat{Y}_k = \hat{m}_h(X_k) = \frac{\sum_{i=1}^{n} K(\frac{X_k - X_i}{h}) R_i Y_i}{\sum_{j=1}^{n} K(\frac{X_k - X_j}{h}) R_j} \tag{4.8}$$

where h is the bandwidth of the kernel K.

The estimation of mean $\bar{Y}$ is given by (3.39):

$$\begin{aligned} \mu_{KR1} &= \frac{1}{n} \sum_{k=1}^{n} \hat{m}_h(X_k) \\ &= \frac{1}{n} \sum_{k=1}^{n} \frac{\sum_{i=1}^{n} K(\frac{X_k - X_i}{h}) R_i Y_i}{\sum_{j=1}^{n} K(\frac{X_k - X_j}{h}) R_j} \end{aligned} \tag{4.9}$$

or by (3.40):

$$\mu_{KR2} = \frac{1}{n} \sum_{i=1}^{n} [R_i Y_i + (1 - R_i) \hat{m}_h(X_i)] \tag{4.10}$$

where $\hat{m}_h(x)$ is given by (4.8).

1. Epanechnikov Kernel

The Kernel is given in (3.20):

$$K(t) = \frac{3}{4}(1 - t^2)I(|t| \leq 1) \tag{4.11}$$

The estimator (4.8) for the Epanechnikov Kernel is illustrated by Fig. 4.4.

Using the estimator (4.9), we obtain:

$$\bar{Y}_{Epanechnikov} \approx 101.2395 \text{ hours}$$

and

$$\text{Error } = |\bar{Y}_{Epanechnikov} - \bar{Y}| \approx 0.6039 \text{ hours}$$
$$\text{Relative error } = |\frac{\bar{Y}_{Epanechnikov} - \bar{Y}}{\bar{Y}}| \approx 0.0059$$

Using the estimator (4.10), we obtain:

$$\bar{Y}_{Epanechnikov} \approx 101.3955 \text{ hours}$$

and

$$\text{Error } = |\bar{Y}_{Epanechnikov} - \bar{Y}| \approx 0.4479 \text{ hours}$$

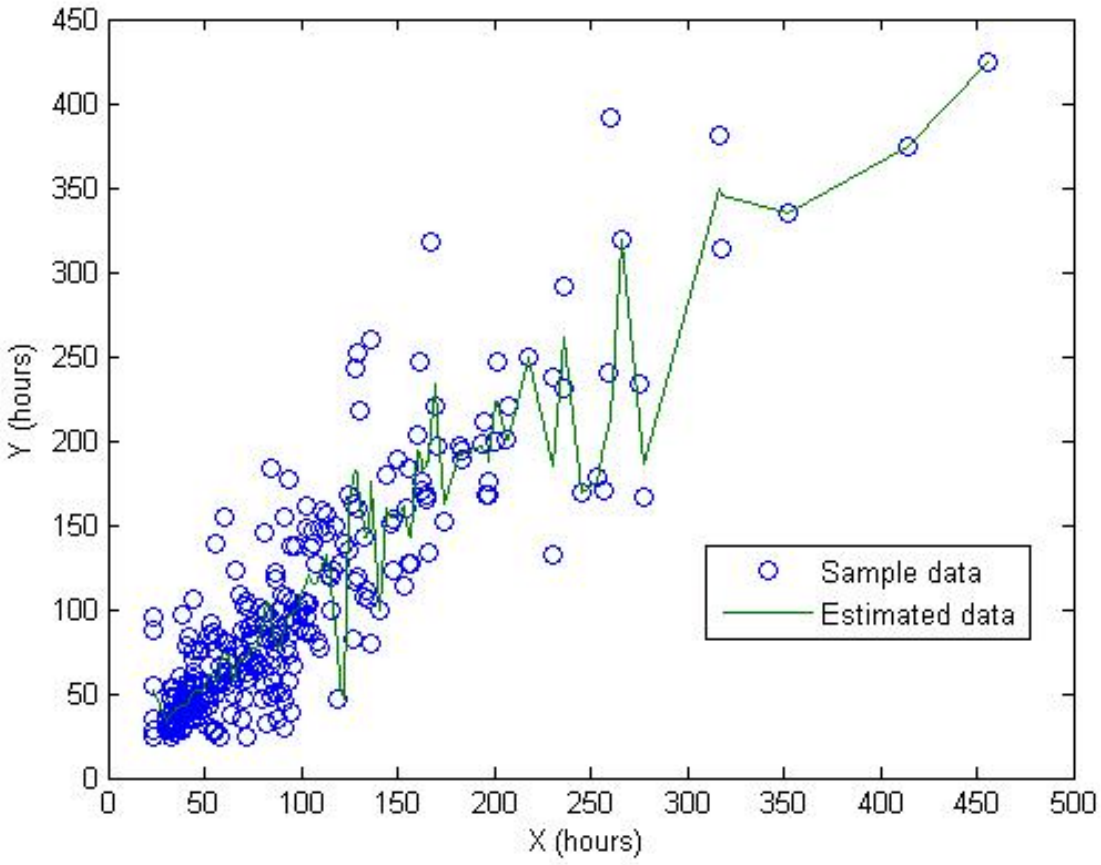

Figure 4.4: Epanechnikov kernel estimator (h = 4)

$$\text{Relative error } = |\frac{\bar{Y}_{Epanechnikov} - \bar{Y}}{\bar{Y}}| \approx 0.0044$$

2. Polynomial Order-4 Kernel

The Kernel is given in (3.20):

$$K(t) = \frac{15}{32}(3 - 10t^2 + 7t^4)I(|t| \leq 1) \qquad (4.12)$$

The estimator (4.8) for the Polynomial Order-4 Kernel is illustrated by Fig. 4.5.

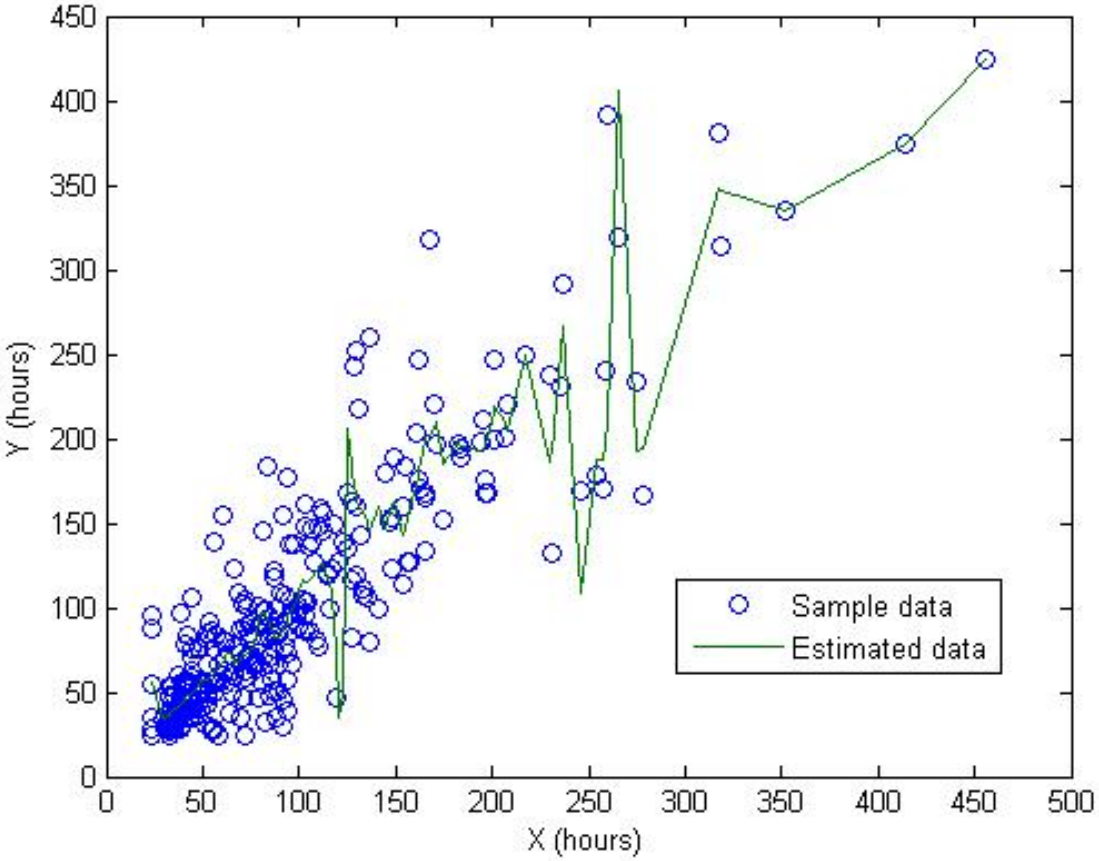

Figure 4.5: Polynomial Order-4 kernel estimator (h = 10)

Using the estimator (4.9), we obtain:

$$\bar{Y}_{PolyOrder4} \approx 101.6475 \text{ hours}$$

and

$$\text{Error } = |\bar{Y}_{PolyOrder4} - \bar{Y}| \approx 0.1959 \text{ hours}$$

$$\text{Relative error } = |\frac{\bar{Y}_{PolyOrder4} - \bar{Y}}{\bar{Y}}| \approx 0.0019$$

Using the estimator (4.10), we obtain:

$$\bar{Y}_{PolyOrder4} \approx 101.3957 \text{ hours}$$

and

$$\text{Error } = |\bar{Y}_{PolyOrder4} - \bar{Y}| \approx 0.4477 \text{ hours}$$
$$\text{Relative error } = |\frac{\bar{Y}_{PolyOrder4} - \bar{Y}}{\bar{Y}}| \approx 0.0044$$

3. Gaussian Kernel

The Kernel is given by (3.21):

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \tag{4.13}$$

The estimator (4.8) for the Gaussian Kernel is illustrated by Fig. 4.6.

Using the estimator (4.9), we obtain:

$$\bar{Y}_{Gauss} \approx 101.7062 \text{ hours}$$

and

$$\text{Error } = |\bar{Y}_{Gauss} - \bar{Y}| \approx 0.1373 \text{ hours}$$

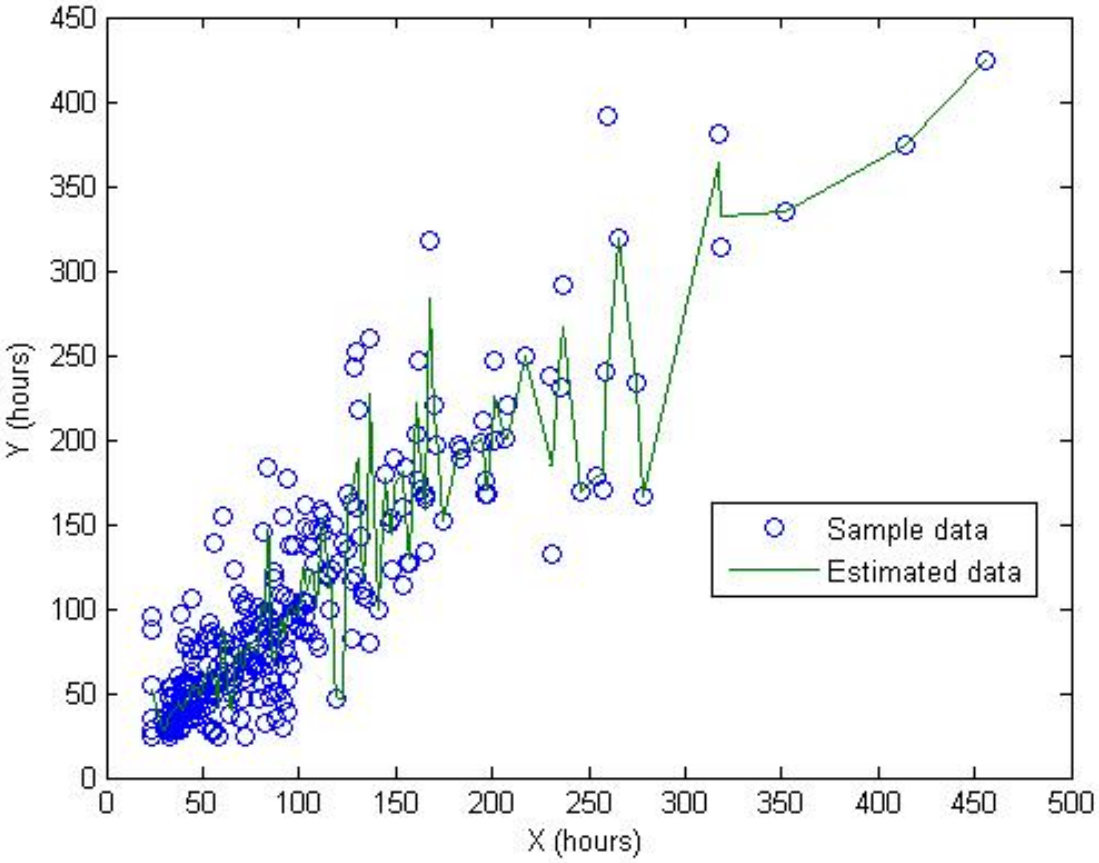

Figure 4.6: Gaussian kernel estimator (h = 0.9)

$$\text{Relative error } = |\frac{\bar{Y}_{Gauss} - \bar{Y}}{\bar{Y}}| \approx 0.0013$$

Using the estimator (4.10), we obtain:

$$\bar{Y}_{Gauss} \approx 101.8267 \text{ hours}$$

and

$$\text{Error } = |\bar{Y}_{Gauss} - \bar{Y}| \approx 0.0167 \text{ hours}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Gauss} - \bar{Y}}{\bar{Y}}| \approx 1.6400e-004 = 0.000164$$

4.2 Multiple Imputation

Here we apply the kernel smoothing methods for multiple imputation described in Section 3.8 to estimate the missing data of sales in the fifth week Y provided that the sample of sales in the first week X_1 and the sample of work hours in the fifth week X_2 are complete.

Let $X_1 s = \{X_{11}, \cdots, X_{n1}\}$ be the sample of sales in week 1, $X2_S = \{X_{12}, \cdots, X_{n2}\}$ be the sample of work hours in week 5, $Y_S = \{Y_1, \cdots, Y_n\}$ be the sample of sales in week 5 for these divisions of the stores, where $n = 287$, and $R = \{R_1, \cdots, R_n\}$ be the response indicator set defined by (2.1). The estimator of Y_j is given by (3.49):

$$\hat{Y}_j = \hat{m}_h(X_{j1}, X_{j2}) = \frac{\sum_{i=1}^{n} K(\frac{X_{j1}-X_{i1}}{h_1}) K(\frac{X_{j2}-X_{i2}}{h_2}) R_i Y_i}{\frac{\sum_{i=1}^{n} R_i}{n} \sum_{i=1}^{n} K(\frac{X_{j1}-X_{i1}}{h_1}) K(\frac{X_{j2}-X_{i2}}{h_2})} \tag{4.14}$$

The estimation of mean $\bar{Y}$ is given by

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n}\sum_{j=1}^{n}\hat{m}_h(X_{j1}, X_{j2}) \\ &= \sum_{j=1}^{n}\frac{\sum_{i=1}^{n} K(\frac{X_{j1}-X_{i1}}{h_1})K(\frac{X_{j2}-X_{i2}}{h_2})R_iY_i}{(\sum_{i=1}^{n} R_i)\sum_{i=1}^{n} K(\frac{X_{j1}-X_{i1}}{h_1})K(\frac{X_{j2}-X_{i2}}{h_2})}\end{aligned} \tag{4.15}$$

or

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}[R_iY_i + (1-R_i)\hat{m}_h(X_{i1}, X_{i2})] \tag{4.16}$$

where $\hat{m}_h(X_{j1}, X_{j2})$ is given by (4.14).

Before we delete some data from the sample Y of sales in the fifth week, the average of Y, given by the data table in Appendix A, is

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \approx 53888 \text{ \$} \tag{4.17}$$

4.2.1 Epanechnikov Kernel

The Kernel is given by (4.11). The estimators of Y_k for the single imputation, $\hat{Y}_k = \hat{m}_{h_1}(X_{k1})$ with $h_1 = 4000$ and $\hat{Y}_k = \hat{m}_{h_2}(X_{k2})$ with $h_2 = 15$, are given by (4.8), that are illustrated by Fig. 4.7 and Fig. 4.8, respectively.

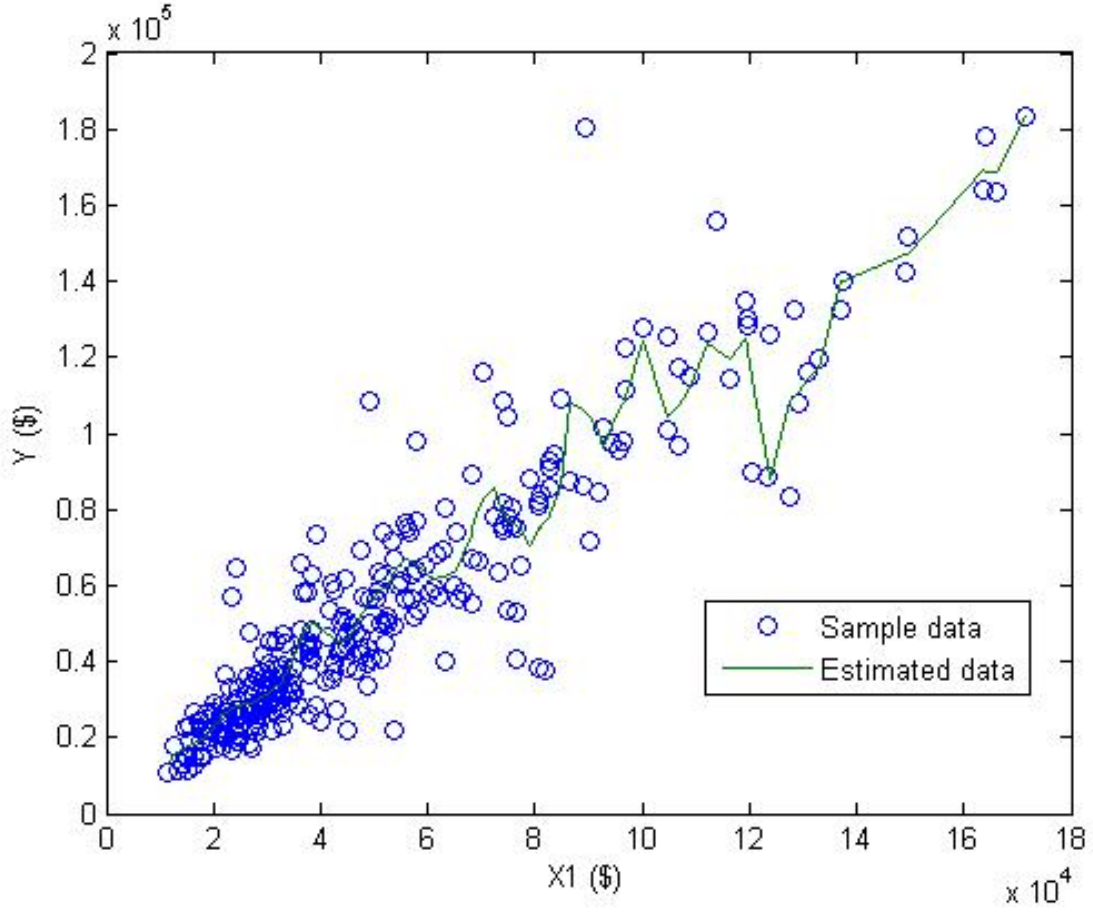

Figure 4.7: Epanechnikov kernel estimator – single imputation by X_1

Using the estimator (4.9), we obtain:

$$\bar{Y}_{Epanechnikov}(X_1) \approx 54744 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Epanechnikov}(X_1) - \bar{Y}| \approx 856.8022 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Epanechnikov}(X_1) - \bar{Y}}{\bar{Y}}| \approx 0.0159$$

and

$$\bar{Y}_{Epanechnikov}(X_2) \approx 54773 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Epanechnikov}(X_2) - \bar{Y}| \approx 885.4971 \text{ \$}$$

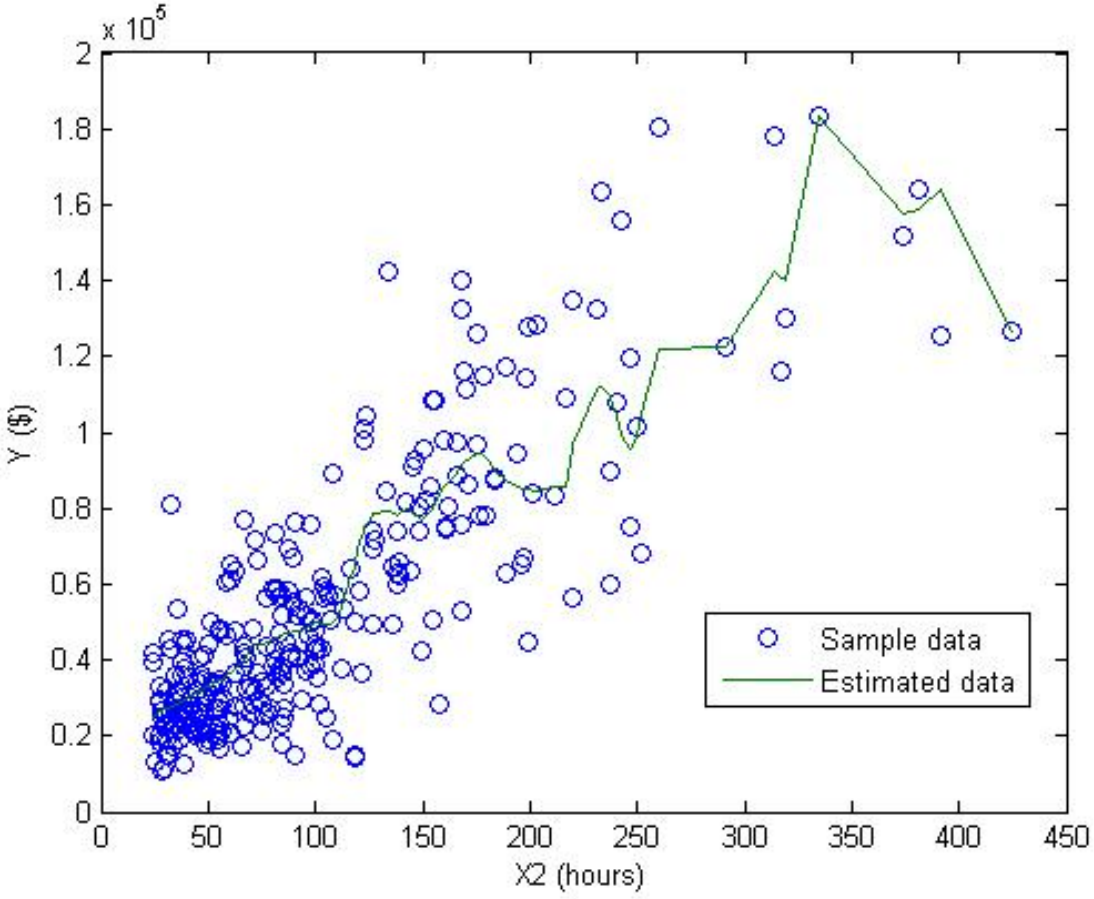

Figure 4.8: Epanechnikov kernel estimator – single imputation by X_2

$$\text{Relative error } = |\frac{\bar{Y}_{Epanechnikov}(X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0164$$

Using the estimator (4.10), we obtain:

$$\bar{Y}_{Epanechnikov}(X_1) \approx 54742 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Epanechnikov}(X_1) - \bar{Y}| \approx 854.4287 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Epanechnikov}(X_1) - \bar{Y}}{\bar{Y}}| \approx 0.0159$$

and

$$\bar{Y}_{Epanechnikov}(X_2) \approx 54902 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Epanechnikov}(X_2) - \bar{Y}| \approx 1014.0 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Epanechnikov}(X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0188$$

By the estimator (4.15) for the multiple imputation method, we obtain:

$$\bar{Y}_{EpanechnikovMul}(X_1, X_2) \approx 57583 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{EpanechnikovMul}(X_1, X_2) - \bar{Y}| \approx 3695.1 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{EpanechnikovMul}(X_1, X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0686$$

4.2.2 Polynomial Order-4 Kernel

The Kernel is given by (4.12). The estimators of Y_k for the single imputation, $\hat{Y}_k = \hat{m}_{h_1}(X_{k1})$ with $h_1 = 10000$ and $\hat{Y}_k = \hat{m}_{h_2}(X_{k2})$ with $h_2 = 20$, are given by (4.8), that are illustrated by Fig. 4.9 and Fig. 4.10, respectively.

Using the estimator (4.9), we obtain:

$$\bar{Y}_{Poly4}(X_1) \approx 54715 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Poly4}(X_1) - \bar{Y}| \approx 826.9704 \text{ \$}$$

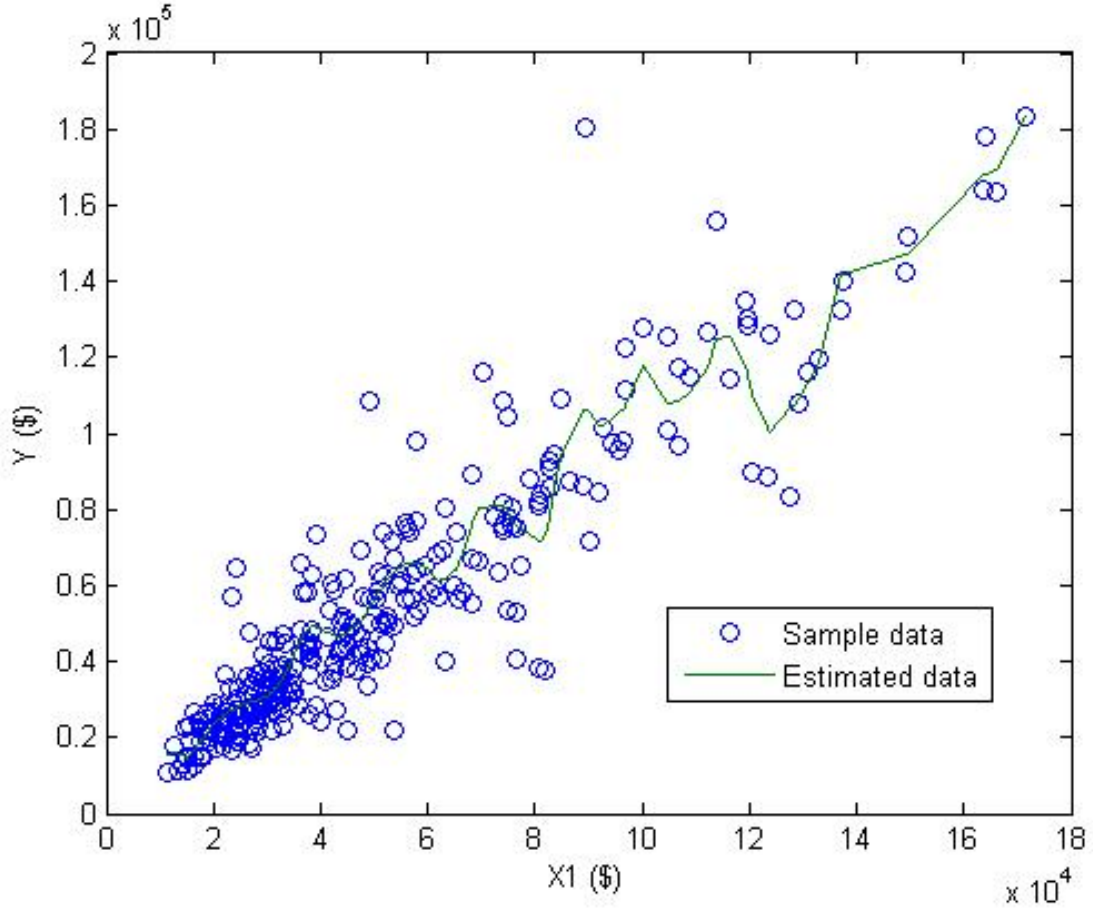

Figure 4.9: Polynomial Order4 Kernel estimator – single imputation by X_1

$$\text{Relative error } = |\frac{\bar{Y}_{Poly4}(X_1) - \bar{Y}}{\bar{Y}}| \approx 0.0153$$

and

$$\bar{Y}_{Poly4}(X_2) \approx 55092 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Poly4}(X_2) - \bar{Y}| \approx 1204.8 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Poly4}(X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0224$$

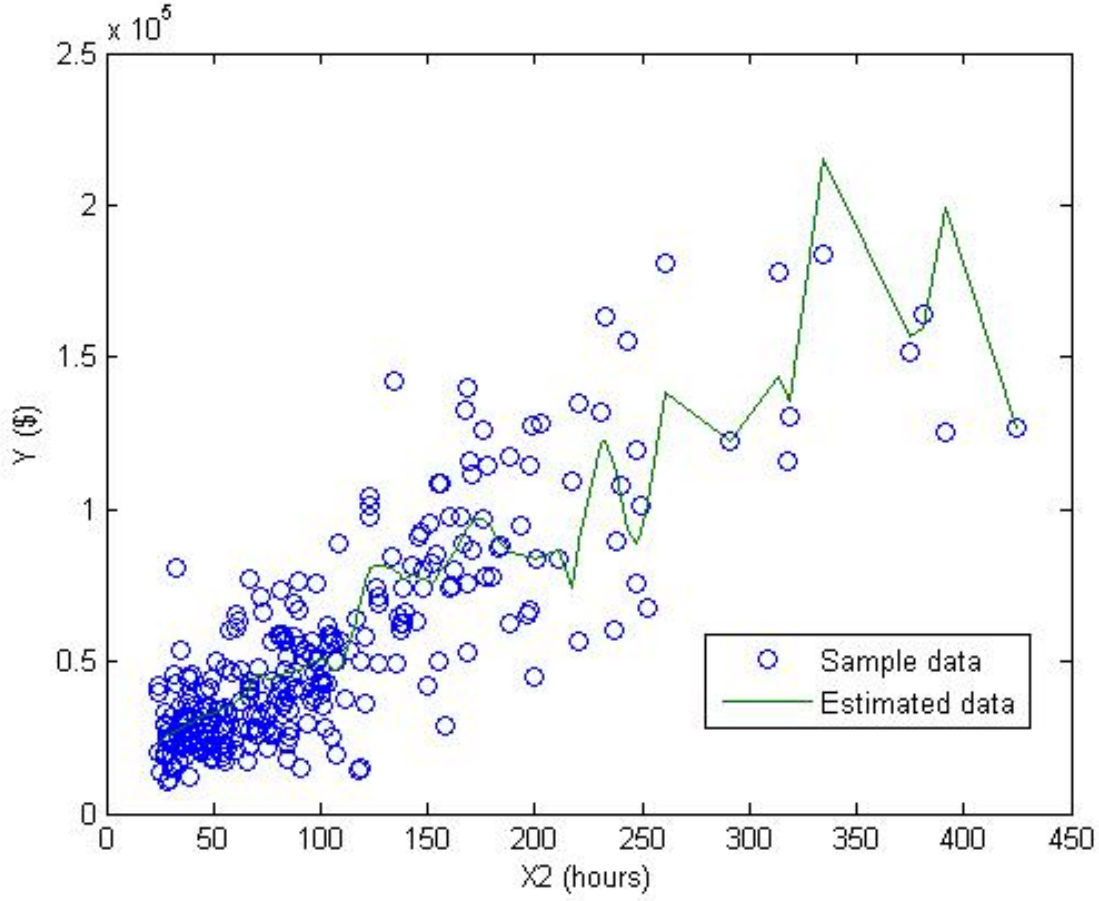

Figure 4.10: Polynomial Order4 Kernel estimator – single imputation by X_2

Using the estimator (4.10), we obtain:

$$\bar{Y}_{Poly4}(X_1) \approx 54767 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Poly4}(X_1) - \bar{Y}| \approx 879.3035 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Poly4}(X_1) - \bar{Y}}{\bar{Y}}| \approx 0.0163$$

and

$$\bar{Y}_{Poly4}(X_2) \approx 55090 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Poly4}(X_2) - \bar{Y}| \approx 1202.1 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Poly4}(X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0223$$

By the estimator (4.15) for the multiple imputation method, we obtain:

$$\bar{Y}_{Poly4Mul}(X_1, X_2) \approx 57735 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Poly4Mul}(X_1, X_2) - \bar{Y}| \approx 3847.0 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Poly4Mul}(X_1, X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0714$$

4.2.3 Gaussian Kernel

The Kernel is given by (4.13). The estimators of Y_k for the single imputation, $\hat{Y}_k = \hat{m}_{h_1}(X_{k1})$ with $h_1 = 100$ and $\hat{Y}_k = \hat{m}_{h_2}(X_{k2})$ with $h_2 = 0.9$, are given by (4.8), that are illustrated by Fig. 4.11 and Fig. 4.12, respectively.

Using the estimator (4.9), we obtain:

$$\bar{Y}_{Gauss}(X_1) \approx 55160 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Gauss}(X_1) - \bar{Y}| \approx 1272.8 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Gauss}(X_1) - \bar{Y}}{\bar{Y}}| \approx 0.0236$$

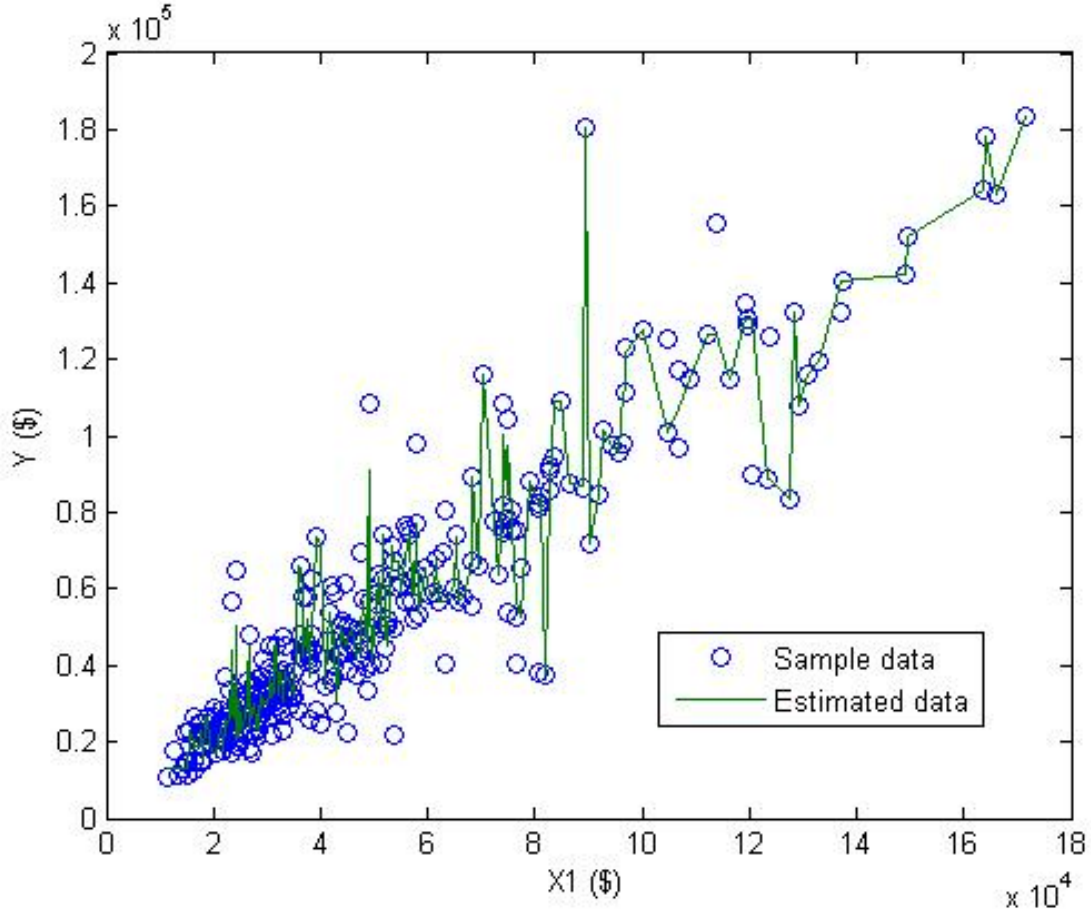

Figure 4.11: Gaussian Kernel estimator – single imputation by X_1

and

$$\bar{Y}_{Gauss}(X_2) \approx 54232 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Gauss}(X_2) - \bar{Y}| \approx 344.1465 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Gauss}(X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0064$$

Using the estimator (4.10), we obtain:

$$\bar{Y}_{Gauss}(X_1) \approx 55170 \text{ \$}$$

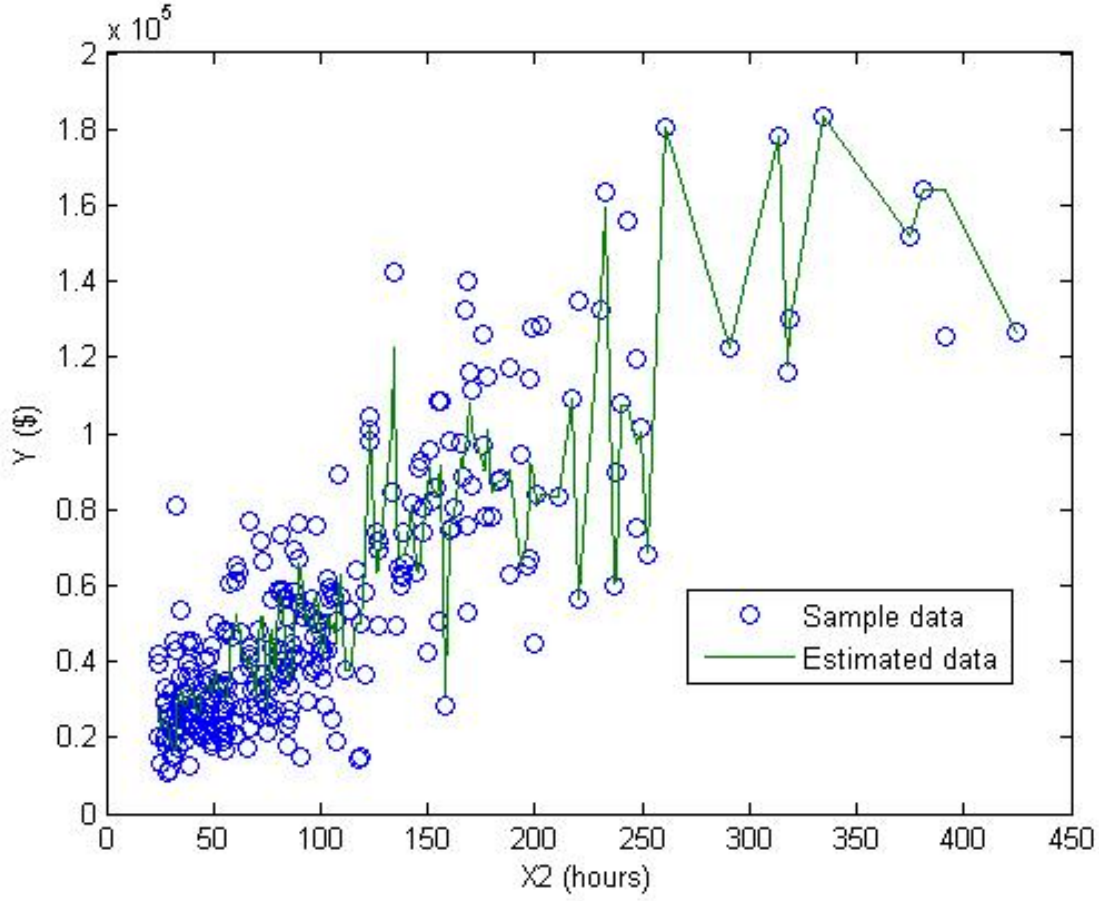

Figure 4.12: Gaussian kernel estimator – single imputation by X_2

$$\text{Error } = |\bar{Y}_{Gauss}(X_1) - \bar{Y}| \approx 1282.4 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Gauss}(X_1) - \bar{Y}}{\bar{Y}}| \approx 0.0238$$

and

$$\bar{Y}_{Gauss}(X_2) \approx 54155 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{Gauss}(X_2) - \bar{Y}| \approx 367.2215 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{Gauss}(X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0050$$

By the estimator (4.15) for the multiple imputation method, we obtain:

$$\bar{Y}_{GaussMul}(X_1, X_2) \approx 57961 \text{ \$}$$

$$\text{Error } = |\bar{Y}_{GaussMul}(X_1, X_2) - \bar{Y}| \approx 4073.5 \text{ \$}$$

$$\text{Relative error } = |\frac{\bar{Y}_{GaussMul}(X_1, X_2) - \bar{Y}}{\bar{Y}}| \approx 0.0756$$

4.3 Summary

The following tables show the results that we obtained in Section 4.1 and Section 4.2. We will compare these mothods and discuss their applicability.

4.3.1 Single Imputation

In Section 4.1, we used some specific methods to estimate the missing data of the work hours in the fifth week by the complete sample of work hours in the first week. The table 4.1 shows the relative errors of those modes that we used in Section 4.1.

We see that, in general, these kernel smoothing methods are better than non-kernel methods.

Modes	**Relative error** (%)
Ratio Method	0.96
Regression Method	0.50
Optimal Method	2.32
Epanechnikov Kernel	0.59
Polynomial Order4 Kernel	0.19
Gaussian Kernel	0.13

Table 4.1: Relative errors of the methods for single imputation

We note, as predicted, that the optimal method is the worst one among these methods in Table 4.1. As mentioned in Section 2.1.5, for estimating the mean $\bar{Y}$, this method needs an another estimation of $\bar{Y}$ to optimize the optimal coefficients A and B.

Table 4.1 also shows that the Gaussian kernel smoothing method is better than those polynomial kernel smoothing methods. Since for the polynomial kernel smoothing methods, the kernel $K(t) = 0$ when $t > 1$, the chosen bandwidth h could not be very small, while for Gaussian kernel, h can be chosen to be enough small to optimize the procedure. This is why the Gaussian kernel smoothing method is the best one among these methods in Table 4.1.

4.3.2 Multiple Imputation

In Section 4.2, we used some specific multiple imputation methods to estimate the missing data of the sales in the fifth week Y by the complete samples of the sales in the first week X_1 and the work hours in the fifth week X_2. The following tables show the relative errors of those modes that we used in Section 4.2.

Modes	**Relative error** (%)
Simple Imputation Y(X_1)	1.59
Simple Imputation Y(X_2)	1.64
Multiple Imputation Y(X_1, X_2)	6.86

Table 4.2: Epanechnikov Kernel

Modes	**Relative error** (%)
Simple Imputation Y(X_1)	1.53
Simple Imputation Y(X_2)	2.24
Multiple Imputation Y(X_1, X_2)	7.14

Table 4.3: Polynomial Order4 Kernel

Modes	**Relative error** (%)
Simple Imputation Y(X_1)	2.36
Simple Imputation Y(X_2)	0.64
Multiple Imputation Y(X_1, X_2)	7.56

Table 4.4: Gaussian Kernel

We see, from Table 4.2, Table 4.3, and Table 4.4, that the multiple imputation methods used here are much worse than the corresponding simple imputation methods. Note that for a multiple imputation kernel smoothing method, the choice of an optimal bandwidth h is very difficult. This might cause the poor accuracy for the multiple imputation kernel smoothing method.

In order to avoid such poor performance of the multiple imputation kernel smoothing method, we can simply use the estimator (2.45) to estimate $\bar{Y}$, i.e.,

$$\hat{\bar{Y}}(X_1, X_2) = \frac{\hat{\bar{Y}}(X_1) + \hat{\bar{Y}}(X_2)}{2} \tag{4.18}$$

This gives

$$\bar{Y}_{EpanechnikovMul}(X_1, X_2) \approx \frac{54744 + 54733}{2} \approx 57583 \text{ \$}$$
$$\text{Error } = |\bar{Y}_{EpanechnikovMul}(X_1, X_2) - \bar{Y}| \approx 870.5 \text{ \$}$$
$$\text{Relative error } = |\frac{\bar{Y}_{EpanechnikovMul}(X_1, X_2) - \bar{Y}}{\bar{Y}}| \approx 1.62\%$$

$$\bar{Y}_{Poly4Mul}(X_1, X_2) \approx \frac{54715 + 55092}{2} \approx 54904 \text{ \$}$$
$$\text{Error } = |\bar{Y}_{Poly4Mul}(X_1, X_2) - \bar{Y}| \approx 1015.5 \text{ \$}$$
$$\text{Relative error } = |\frac{\bar{Y}_{Poly4Mul}(X_1, X_2) - \bar{Y}}{\bar{Y}}| \approx 1.88\%$$

$$\bar{Y}_{GaussMul}(X_1, X_2) \approx \frac{55160 + 54232}{2} \approx 54696 \text{ \$}$$
$$\text{Error } = |\bar{Y}_{GaussMul}(X_1, X_2) - \bar{Y}| \approx 808.0 \text{ \$}$$
$$\text{Relative error } = |\frac{\bar{Y}_{GaussMul}(X_1, X_2) - \bar{Y}}{\bar{Y}}| \approx 1.50\%$$

Then the situation is much more improved as shown in Table 4.5.

Modes	**Relative error** (%)
Epanechnikov Kernel	1.62
Polynomial Order4 Kernel	1.88
Gaussian Kernel	1.50

Table 4.5: Relative errors of the methods for multiple imputation

Furthermore, it verifies the conclusion in the section 4.3.1, i.e., the Gaussian kernel smoothing method is the best one among these methods listed here.

Chapter 5

Conclusion

In this thesis, we reviewed several single imputation and multiple imputation techniques to deal with the problem of missing data in a census or a sample survey. Also we present some estimators for the missing data and the theories about the variance and mean squared error of those estimators. Finally, with some examples, we compare those modes to find their advantages, disadvantages, and applicabilities.

For a single imputation problem, the estimator of an element Y_k from the sample Y_S is given by a function $\hat{m}$ of the sample X_S:

$$\hat{Y_k} = \hat{m}_k(X_S) \tag{5.1}$$

Usually, there are two classifications for the estimation functions $\hat{m}$: single depended and linear: $\hat{Y_k} = A + BX_k$, multiple depended and nonlinear: $\hat{Y_k} = \hat{m}(k, X_S)$.

Those single depended and linear functions $\hat{m}$ are introduced in Chapter 2. We present and discuss in Section 2.1 the mean method, the ratio method, the regression method, the power transformation method, and the optimal method for imputation.

Those multiple depended and nonlinear functions $\hat{m}$ are introduced in Chapter 3. We present and discuss in Chapter 3 the kernel smoothing method, the k-nearest neighbor method, etc. We introduce some theories for the selection of kernel function and bandwidth. The application in Chapter 4 shows how important about the choice of the bandwidth h.

For a multiple imputation problem, the estimator of an element Y_k is also represented by (5.1), where every element X_k in the sample X_S is a real vector in stead of a real number.

In Section 2.2, we introduce the estimator (2.45) of the multiple imputation problem: an average of all single imputations. In Section 3.8, we introduce another estimator (3.49): the multiple kernel smoothing. The application in Chapter 4 shows that the estimator (2.45) is simple and applicable, while the estimator (3.49) is complicated and needs more works to optimize the bandwidth h.

In Chapter 4, we apply these modes introduced in Chapter 2 and Chapter 3 to a real case. We calculate the estimators by single imputation techniques and multiple imputation techniques, then compare those imputation methods and conclude their advantages, disadvantages, and applicabilities.

In practice, all of these imputation techniques used in this thesis work well for the missing data problem. For the case of Chapter 4, those kernel smoothing methods are better than the linear imputation methods. According to results of the application, the Gaussian kernel smoothing method is a very good approach to the missing data problem.

In general, every specific missing data problem has a most appropriate specific approach for it. Choosing a proper mode is very important to resolve the problem. For a kernel smoothing method, the choice of an optimal bandwidth h tends to become the critical to success.

Appendix A

Data Table

The following tables contain the data using for the application in Chapter 4. The data, given by Professor Wei Sun of Department of Mathematics and Statistics of Concordia University, are quoted from the author's report of the B Sc.Honors project in 2011.

WK	StoreID	Division	Position	Hours	Sales
1	NO.100	NO.1	1	62.14	27352.78
1	NO.101	NO.1	1	105.64	24165.36
1	NO.102	NO.1	1	109.88	48734.1
1	NO.104	NO.1	1	58.44	36891.84
1	NO.105	NO.1	1	64.06	18891.34
1	NO.106	NO.1	1	148.16	104730.56
1	NO.107	NO.1	1	94.92	38221.5
1	NO.108	NO.1	1	38.86	35088.12
1	NO.109	NO.1	1	132.46	73923.76
1	NO.110	NO.1	1	44.32	27820.52
1	NO.1102	NO.11	1	93.36	37861.6
1	NO.1105	NO.11	1	39.58	16296.1
1	NO.1107	NO.11	1	34.16	30758.76
1	NO.1108	NO.11	1	76.56	20401.82
1	NO.1109	NO.11	1	40.26	27017.52
1	NO.111	NO.1	1	47.74	26338.34
1	NO.1111	NO.11	1	111.12	27671.66
1	NO.1116	NO.11	1	33.16	14114.66
1	NO.112	NO.1	1	52.36	23706.9
1	NO.1128	NO.11	1	75.88	38208.08
1	NO.113	NO.1	1	45.08	30696.22
1	NO.115	NO.1	1	147.68	82638.1
1	NO.116	NO.1	1	52.52	29069.18
1	NO.117	NO.1	1	32.76	24627.08
1	NO.118	NO.1	1	90.66	62205.72
1	NO.119	NO.1	1	318.22	164196.9
1	NO.120	NO.1	1	52.94	39345.94
1	NO.1200	NO.12	1	49.12	31673.1
1	NO.1201	NO.12	1	44.72	42117.84
1	NO.1202	NO.12	1	102.26	50420.74

WK	StoreID	Division	Position	WorkHours	Sales
5	NO.100	NO.1	1	69.96	26087.2
5	NO.101	NO.1	1	136	64564.56
5	NO.102	NO.1	1	77.18	43695.44
5	NO.104	NO.1	1	24	41985.58
5	NO.105	NO.1	1	37.86	26394.12
5	NO.106	NO.1	1	122.6	101016.8
5	NO.107	NO.1	1	137.9	62778.94
5	NO.108	NO.1	1	35.46	31842.12
5	NO.109	NO.1	1	142.18	81460.38
5	NO.110	NO.1	1	36.7	25131.4
5	NO.1102	NO.11	1	73.02	36836.14
5	NO.1105	NO.11	1	38.4	12288.28
5	NO.1107	NO.11	1	53.04	27212.42
5	NO.1108	NO.11	1	65.7	17027.42
5	NO.1109	NO.11	1	55.48	23183.44
5	NO.111	NO.1	1	36.7	31358.84
5	NO.1111	NO.11	1	158.14	28511.16
5	NO.1116	NO.11	1	25	13263.88
5	NO.112	NO.1	1	42.14	21683.58
5	NO.1128	NO.11	1	91.4	40575.78
5	NO.113	NO.1	1	44.82	31619.38
5	NO.115	NO.1	1	153.46	85506.74
5	NO.116	NO.1	1	66.64	27768.34
5	NO.117	NO.1	1	30	25251.84
5	NO.118	NO.1	1	109	56920.64
5	NO.119	NO.1	1	313.84	178066.4
5	NO.120	NO.1	1	81.52	73258.68
5	NO.1200	NO.12	1	39.62	45328.2
5	NO.1201	NO.12	1	58.04	60597.76
5	NO.1202	NO.12	1	104.06	56157.5

WK	StoreID	Division	Position	Hours	Sales
1	NO.205	NO.2	1	69.72	49607.16
1	NO.206	NO.2	1	316.94	163457.24
1	NO.207	NO.2	1	55.84	24070.58
1	NO.208	NO.2	1	164.92	76618.24
1	NO.209	NO.2	1	193.62	116461.46
1	NO.210	NO.2	1	140.72	46679.4
1	NO.211	NO.2	1	60.48	49140.72
1	NO.212	NO.2	1	183.5	106927.6
1	NO.213	NO.2	1	127.4	75725.1
1	NO.214	NO.2	1	414.66	149549.42
1	NO.215	NO.2	1	32.88	34304.14
1	NO.216	NO.2	1	74.7	44061.6
1	NO.217	NO.2	1	146.78	95798.74
1	NO.218	NO.2	1	106.42	65122.88
1	NO.219	NO.2	1	259	129238.66
1	NO.220	NO.2	1	102.62	76516.58
1	NO.221	NO.2	1	81.94	43692.7
1	NO.222	NO.2	1	165.32	94510.36
1	NO.223	NO.2	1	112.74	74053.7
1	NO.224	NO.2	1	114.9	60274.32
1	NO.225	NO.2	1	162.18	106929.02
1	NO.226	NO.2	1	81	75896.28
1	NO.227	NO.2	1	79.12	47296.52
1	NO.228	NO.2	1	88.44	47865.96
1	NO.229	NO.2	1	82.04	43951.62
1	NO.230	NO.2	1	201.6	132899
1	NO.231	NO.2	1	69.82	69649.68
1	NO.232	NO.2	1	84.1	78869.38
1	NO.233	NO.2	1	165.56	148930.6
1	NO.234	NO.2	1	155.2	86659.5

WK	StoreID	Division	Position	Hours	Sales
5	NO.205	NO.2	1	88.06	40393.84
5	NO.206	NO.2	1	381.24	164293
5	NO.207	NO.2	1	49.58	29880.6
5	NO.208	NO.2	1	167.98	52727.78
5	NO.209	NO.2	1	197.74	114597.4
5	NO.210	NO.2	1	99.92	47830.6
5	NO.211	NO.2	1	155.1	108344.1
5	NO.212	NO.2	1	188.66	117404.5
5	NO.213	NO.2	1	161.92	80235.68
5	NO.214	NO.2	1	374.32	151972.2
5	NO.215	NO.2	1	47.42	35695.06
5	NO.216	NO.2	1	84.64	51605.46
5	NO.217	NO.2	1	150.8	95711.34
5	NO.218	NO.2	1	138.56	73986.72
5	NO.219	NO.2	1	240	107878.2
5	NO.220	NO.2	1	161.34	75146.12
5	NO.221	NO.2	1	98.62	50562.5
5	NO.222	NO.2	1	165.56	97296.36
5	NO.223	NO.2	1	155.74	108285.9
5	NO.224	NO.2	1	120.94	58193.1
5	NO.225	NO.2	1	175.76	96789.52
5	NO.226	NO.2	1	97.98	75499.1
5	NO.227	NO.2	1	100.58	43074.86
5	NO.228	NO.2	1	83.98	57041.9
5	NO.229	NO.2	1	81.7	42706.86
5	NO.230	NO.2	1	247	119354.3
5	NO.231	NO.2	1	72.7	66093.68
5	NO.232	NO.2	1	183.98	87827.4
5	NO.233	NO.2	1	133.96	142313.3
5	NO.234	NO.2	1	183.34	87329.24

WK	StoreID	Division	Position	Hours	Sales	WK	StoreID	Division	Position	WorkHours	Sales	WK	StoreID	Division	Position	Hours	Sales	WK	StoreID	Division	Position	Hours	Sales
1	NO.1203	NO.12	1	91.12	51990.42	5	NO.1203	NO.12	1	154.96	50409.6	1	NO.235	NO.2	1	74.4	32919.36	5	NO.235	NO.2	1	67.66	32689.18
1	NO.1206	NO.12	1	37.56	30412.34	5	NO.1206	NO.12	1	41.26	35638.3	1	NO.236	NO.2	1	44.72	23513.96	5	NO.236	NO.2	1	105.8	56815.04
1	NO.1208	NO.12	1	78.9	58058.64	5	NO.1208	NO.12	1	66.36	76816.9	1	NO.238	NO.2	1	103.44	65911.64	5	NO.238	NO.2	1	95.62	56682.38
1	NO.1209	NO.12	1	39.66	30440.42	5	NO.1209	NO.12	1	37.86	30367.1	1	NO.239	NO.2	1	136.54	89347.2	5	NO.239	NO.2	1	260.32	180632.7
1	NO.121	NO.1	1	99.88	53690.34	5	NO.121	NO.1	1	89.2	67087.1	1	NO.241	NO.2	1	149.62	57111.38	5	NO.241	NO.2	1	188.62	62715.96
1	NO.1211	NO.12	1	32.76	21640.42	5	NO.1211	NO.12	1	27.78	20246.4	1	NO.242	NO.2	1	169.94	55906.54	5	NO.242	NO.2	1	220.34	56574.84
1	NO.1212	NO.12	1	36.02	30615.08	5	NO.1212	NO.12	1	38.1	45307.42	1	NO.243	NO.2	1	133.46	46607.46	5	NO.243	NO.2	1	111.84	37584.56
1	NO.1213	NO.12	1	61.3	50666.96	5	NO.1213	NO.12	1	62.18	63558.38	1	NO.244	NO.2	1	88.28	57139.42	5	NO.244	NO.2	1	83.36	56470.08
1	NO.1214	NO.12	1	42.98	16210.38	5	NO.1214	NO.12	1	35	26462.5	1	NO.245	NO.2	1	99.18	42202.9	5	NO.245	NO.2	1	101.06	35479.18
1	NO.1215	NO.12	1	87.32	23066.22	5	NO.1215	NO.12	1	50.88	33288.74	1	NO.246	NO.2	1	266.1	119479.1	5	NO.246	NO.2	1	319.26	130298.2
1	NO.1216	NO.12	1	42.92	38392.52	5	NO.1216	NO.12	1	49.74	44075.38	1	NO.248	NO.2	1	45.94	17628.5	5	NO.248	NO.2	1	54.74	18878.34
1	NO.1218	NO.12	1	54.78	32166.52	5	NO.1218	NO.12	1	27.4	29404.48	1	NO.249	NO.2	1	193.6	100140.36	5	NO.249	NO.2	1	198.68	127652.6
1	NO.123	NO.1	1	95.76	58336.92	5	NO.123	NO.1	1	93	53434.96	1	NO.250	NO.2	1	206.72	81064.68	5	NO.250	NO.2	1	200.74	83992.92
1	NO.124	NO.1	1	89.42	32275.18	5	NO.124	NO.1	1	84.6	38168.94	1	NO.251	NO.2	1	455.6	112023.76	5	NO.251	NO.2	1	424.7	126562.3
1	NO.125	NO.1	1	55.72	29375.34	5	NO.125	NO.1	1	53.96	33650.96	1	NO.252	NO.2	1	127.22	63054.2	5	NO.252	NO.2	1	81.7	40270.78
1	NO.127	NO.1	1	71.06	27336.1	5	NO.127	NO.1	1	46.88	27991.44	1	NO.253	NO.2	1	125.12	53818.58	5	NO.253	NO.2	1	135.52	49528.02
1	NO.128	NO.1	1	196	137596.96	5	NO.128	NO.1	1	168.22	140249.7	1	NO.254	NO.2	1	43.18	19308.46	5	NO.254	NO.2	1	43.6	20229.22
1	NO.129	NO.1	1	94.28	81757.42	5	NO.129	NO.1	1	38.82	37493.62	1	NO.255	NO.2	1	73.36	32789.56	5	NO.255	NO.2	1	100.12	38874.08
1	NO.130	NO.1	1	161.72	56065.2	5	NO.130	NO.1	1	246.9	75359.1	1	NO.256	NO.2	1	129.36	74040.98	5	NO.256	NO.2	1	160.48	74489.62
1	NO.1301	NO.13	1	71.92	47576	5	NO.1301	NO.13	1	58.7	46726.38	1	NO.257	NO.2	1	66.1	58057.7	5	NO.257	NO.2	1	122.68	97801.66
1	NO.1302	NO.13	1	162.78	89003.58	5	NO.1302	NO.13	1	170.76	86546.98	1	NO.258	NO.2	1	183.58	83741.48	5	NO.258	NO.2	1	193.6	94452.6
1	NO.1303	NO.13	1	144.8	72530.16	5	NO.1303	NO.13	1	179.58	77805.84	1	NO.259	NO.2	1	40.92	38001.84	5	NO.259	NO.2	1	78.2	25996.22
1	NO.1305	NO.13	1	92.56	49312.18	5	NO.1305	NO.13	1	76.94	56647.76	1	NO.260	NO.2	1	106.18	82726.12	5	NO.260	NO.2	1	146.4	92402.14
1	NO.1306	NO.13	1	33.82	25802.92	5	NO.1306	NO.13	1	38.54	26574.56	1	NO.261	NO.2	1	120.48	62864.74	5	NO.261	NO.2	1	127.02	69375.46
1	NO.1311	NO.13	1	105.64	52057.68	5	NO.1311	NO.13	1	86.3	44630.14	1	NO.262	NO.2	1	39.28	26426.86	5	NO.262	NO.2	1	37.4	35835.84
1	NO.1312	NO.13	1	59.62	43114.16	5	NO.1312	NO.13	1	55.7	27475.96	1	NO.263	NO.2	1	40.6	21992.82	5	NO.263	NO.2	1	44.58	20907.92
1	NO.1313	NO.13	1	85.92	30087.4	5	NO.1313	NO.13	1	57.48	29738.5	1	NO.264	NO.2	1	60.42	77419.62	5	NO.264	NO.2	1	60.26	65222.12
1	NO.1314	NO.13	1	38.34	15337.66	5	NO.1314	NO.13	1	29.88	11496.84	1	NO.265	NO.2	1	230	120653.38	5	NO.265	NO.2	1	237.5	89600.98
1	NO.1315	NO.13	1	194.8	127675.12	5	NO.1315	NO.13	1	211.52	83589.38	1	NO.266	NO.2	1	352.28	171422.92	5	NO.266	NO.2	1	334.56	183615
1	NO.1319	NO.13	1	37.66	29088.7	5	NO.1319	NO.13	1	34.9	30536.56	1	NO.267	NO.2	1	154.36	96480.9	5	NO.267	NO.2	1	160.28	97684.7

WK	StoreID	Division	Position	Hours	Sales	WK	StoreID	Division	Position	WorkHours	Sales	WK	StoreID	Division	Position	Hours	Sales	WK	StoreID	Division	Position	Hours	Sales
1	NO.1320	NO.13	1	71.48	22403	5	NO.1320	NO.13	1	62.94	27798.06	1	NO.268	NO.2	1	196.6	123913.42	5	NO.268	NO.2	1	175.84	126100.2
1	NO.1321	NO.13	1	72.22	61125.24	5	NO.1321	NO.13	1	104.16	59143.7	1	NO.270	NO.2	1	207.5	119086.86	5	NO.270	NO.2	1	220.08	134686.6
1	NO.1322	NO.13	1	33.08	27090.04	5	NO.1322	NO.13	1	52.7	33791.2	1	NO.271	NO.2	1	33	29007.46	5	NO.271	NO.2	1	46.24	29814.14
1	NO.1325	NO.13	1	49.1	27808.96	5	NO.1325	NO.13	1	56.08	21420.62	1	NO.272	NO.2	1	197.64	137093.32	5	NO.272	NO.2	1	167.86	132425.9
1	NO.1326	NO.13	1	88.52	43276.24	5	NO.1326	NO.13	1	99.84	42646.4	1	NO.273	NO.2	1	44.76	34228.44	5	NO.273	NO.2	1	70.06	32856.9
1	NO.1327	NO.13	1	230.28	64849.46	5	NO.1327	NO.13	1	237.2	60194.08	1	NO.274	NO.2	1	122.76	51514.22	5	NO.274	NO.2	1	138.94	62449.18
1	NO.1328	NO.13	1	55.34	33040.2	5	NO.1328	NO.13	1	28.46	26827.02	1	NO.275	NO.2	1	42.1	21173.64	5	NO.275	NO.2	1	47.14	20792.32
1	NO.1329	NO.13	1	48.58	24766.18	5	NO.1329	NO.13	1	54.68	20019.8	1	NO.276	NO.2	1	112.88	82673.84	5	NO.276	NO.2	1	145.3	91011.02
1	NO.133	NO.1	1	75.84	33414.82	5	NO.133	NO.1	1	71.9	32004.46	1	NO.277	NO.2	1	24	47583.88	5	NO.277	NO.2	1	87.76	69401.02
1	NO.1330	NO.13	1	55.36	36265.66	5	NO.1330	NO.13	1	139.38	65841.56	1	NO.278	NO.2	1	259.82	104655.44	5	NO.278	NO.2	1	391.82	125397.9
1	NO.1331	NO.13	1	24	28721.5	5	NO.1331	NO.13	1	96	36980.54	1	NO.279	NO.2	1	127.68	57743.38	5	NO.279	NO.2	1	116.64	63823.04
1	NO.1332	NO.13	1	38.1	41305.4	5	NO.1332	NO.13	1	96.74	46306.54	1	NO.281	NO.2	1	127.92	113740.42	5	NO.281	NO.2	1	243.04	155579.7
1	NO.1333	NO.13	1	170.82	59405.06	5	NO.1333	NO.13	1	196.22	65198.64	1	NO.282	NO.2	1	125.04	73848.38	5	NO.282	NO.2	1	168.32	75991.04
1	NO.1334	NO.13	1	94.24	42665.72	5	NO.1334	NO.13	1	83.42	36869.58	1	NO.283	NO.2	1	82.54	80711.64	5	NO.283	NO.2	1	32.74	80918.16
1	NO.1336	NO.13	1	39.56	15311.76	5	NO.1336	NO.13	1	31.28	15114.86	1	NO.284	NO.2	1	86.98	31468	5	NO.284	NO.2	1	75.24	34249.06
1	NO.1338	NO.13	1	41.2	28906.52	5	NO.1338	NO.13	1	55.74	28764.36	1	NO.285	NO.2	1	40.08	39044.86	5	NO.285	NO.2	1	42.04	28401.04
1	NO.134	NO.1	1	278.08	123303.52	5	NO.134	NO.1	1	166.16	88804.96	1	NO.287	NO.2	1	136.78	80657.74	5	NO.287	NO.2	1	79.74	38331.9
1	NO.1342	NO.13	1	80.78	73257.24	5	NO.1342	NO.13	1	144.8	63526.64	1	NO.289	NO.2	1	84.92	34174.22	5	NO.289	NO.2	1	93.44	29623.84
1	NO.1345	NO.13	1	34	18017.02	5	NO.1345	NO.13	1	30.08	14683.44	1	NO.290	NO.2	1	51.18	37472.8	5	NO.290	NO.2	1	31.44	45423.14
1	NO.1347	NO.13	1	30.18	24815.78	5	NO.1347	NO.13	1	27.74	33003.14	1	NO.291	NO.2	1	160.5	119689.16	5	NO.291	NO.2	1	202.82	128493.6
1	NO.1350	NO.13	1	74.74	90297.36	5	NO.1350	NO.13	1	71.96	71605.66	1	NO.292	NO.2	1	86.94	32529.68	5	NO.292	NO.2	1	121.24	36396.44
1	NO.1351	NO.13	1	129.26	51020.84	5	NO.1351	NO.13	1	118.8	49754.62	1	NO.293	NO.2	1	81.46	45104.72	5	NO.293	NO.2	1	54.46	22187.04
1	NO.1352	NO.13	1	132.84	19317.22	5	NO.1352	NO.13	1	107.76	19230.72	1	NO.294	NO.2	1	91.76	53893.1	5	NO.294	NO.2	1	45.76	21907.2
1	NO.1353	NO.13	1	48.3	24408.44	5	NO.1353	NO.13	1	49.72	18897.98	1	NO.297	NO.2	1	94.74	57479.8	5	NO.297	NO.2	1	95.68	51964.7
1	NO.1357	NO.13	1	55.12	26801.24	5	NO.1357	NO.13	1	27.82	18433.18	1	NO.298	NO.2	1	118.38	36312.72	5	NO.298	NO.2	1	149.44	42126
1	NO.1358	NO.13	1	45.58	24540.74	5	NO.1358	NO.13	1	40.02	22582.64	1	NO.299	NO.2	1	153.36	74741.06	5	NO.299	NO.2	1	113.5	53710.06
1	NO.1363	NO.13	1	157.06	51518.94	5	NO.1363	NO.13	1	126.42	73880.38	1	NO.407	NO.4	1	84.7	34958.46	5	NO.407	NO.4	1	66.98	38311.56
1	NO.1364	NO.13	1	87.9	25161.18	5	NO.1364	NO.13	1	34.82	23047.34	1	NO.512	NO.5	1	115.06	16255.36	5	NO.512	NO.5	1	118.48	15061.04
1	NO.1366	NO.13	1	61.5	29211.54	5	NO.1366	NO.13	1	79.12	27440.38	1	NO.559	NO.5	1	24.04	11310	5	NO.559	NO.5	1	28.3	10819.08
1	NO.1367	NO.13	1	24	22498.5	5	NO.1367	NO.13	1	24	20124.44	1	NO.705	NO.7	1	64.6	14521.9	5	NO.705	NO.7	1	49.18	22557.02

WK	StoreID	Division	Position	Hours	Sales	WK	StoreID	Division	Position	WorkHours	Sales	WK	StoreID	Division	Position	Hours	Sales	WK	StoreID	Division	Position	Hours	Sales
1	NO.1369	NO.13	1	85.56	28591.8	5	NO.1369	NO.13	1	84.58	23082.3	1	NO.709	NO.7	1	52.54	12525.64	5	NO.709	NO.7	1	49.16	17780.68
1	NO.137	NO.1	1	81.46	55925.78	5	NO.137	NO.1	1	90.18	76242.58	1	NO.727	NO.7	1	36	17493.96	5	NO.727	NO.7	1	32.2	21932.64
1	NO.1370	NO.13	1	24	26566.76	5	NO.1370	NO.13	1	55	47873.86	1	NO.754	NO.7	1	33.34	20105.84	5	NO.754	NO.7	1	27.38	28966.04
1	NO.1371	NO.13	1	43.74	38373.96	5	NO.1371	NO.13	1	55.46	48145.96	1	NO.770	NO.7	1	70.52	21198.66	5	NO.770	NO.7	1	102.4	28321.62
1	NO.1376	NO.13	1	42.56	23369.16	5	NO.1376	NO.13	1	55.34	16905.94	1	NO.771	NO.7	1	72.3	42626.98	5	NO.771	NO.7	1	24	39744.58
1	NO.1377	NO.13	1	62.02	40946.08	5	NO.1377	NO.13	1	55.16	34693.02	1	NO.772	NO.7	1	36.06	15744.72	5	NO.772	NO.7	1	30.98	19021.5
1	NO.138	NO.1	1	109.94	56658.6	5	NO.138	NO.1	1	148.14	74100.5	1	NO.773	NO.7	1	77.96	51063.7	5	NO.773	NO.7	1	67.36	40514.02
1	NO.1381	NO.13	1	94.68	22394.08	5	NO.1381	NO.13	1	104.82	25091.22	1	NO.775	NO.7	1	95.92	44065.4	5	NO.775	NO.7	1	67.1	43429.24
1	NO.1383	NO.13	1	52.28	21915.3	5	NO.1383	NO.13	1	84.68	17880.34	1	NO.776	NO.7	1	65.76	54869.04	5	NO.776	NO.7	1	60.5	61093.36
1	NO.1389	NO.13	1	38.36	24556.28	5	NO.1389	NO.13	1	36.6	18842.4	1	NO.779	NO.7	1	84.38	29042.88	5	NO.779	NO.7	1	82.44	35353.04
1	NO.139	NO.1	1	108.66	66700.68	5	NO.139	NO.1	1	81.32	57970.48	1	NO.780	NO.7	1	103.44	22465.84	5	NO.780	NO.7	1	85.86	24935.56
1	NO.1393	NO.13	1	52.5	26954.78	5	NO.1393	NO.13	1	46.12	26685.54	1	NO.781	NO.7	1	30.36	21249.56	5	NO.781	NO.7	1	36.26	25803.08
1	NO.1396	NO.13	1	54.28	26875.78	5	NO.1396	NO.13	1	85.62	27177.42	1	NO.782	NO.7	1	29.9	43673.9	5	NO.782	NO.7	1	32.94	42997.74
1	NO.1397	NO.13	1	43.04	17779.56	5	NO.1397	NO.13	1	38.2	24474.74	1	NO.785	NO.7	1	99.26	45153.42	5	NO.785	NO.7	1	97.4	38558.88
1	NO.143	NO.1	1	38.9	22898.74	5	NO.143	NO.1	1	28.4	24589.42	1	NO.787	NO.7	1	37.8	24454.74	5	NO.787	NO.7	1	38.78	27263.4
1	NO.144	NO.1	1	36.92	22473.74	5	NO.144	NO.1	1	42.94	26330.42	1	NO.789	NO.7	1	44.18	27860.08	5	NO.789	NO.7	1	63.08	36622.92
1	NO.145	NO.1	1	66.8	44898.72	5	NO.145	NO.1	1	62.24	47369	1	NO.792	NO.7	1	97.2	54493.3	5	NO.792	NO.7	1	137.64	60203.42
1	NO.149	NO.1	1	41.3	21529.42	5	NO.149	NO.1	1	38.28	24233.34	1	NO.798	NO.7	1	57.98	21317.86	5	NO.798	NO.7	1	66.5	22441.28
1	NO.151	NO.1	1	91.62	15619.9	5	NO.151	NO.1	1	29.82	22828.46	1	NO.802	NO.8	1	57	22232.96	5	NO.802	NO.8	1	83.84	36741.32
1	NO.152	NO.1	1	45.52	33129.12	5	NO.152	NO.1	1	34.8	35687.56	1	NO.804	NO.8	1	102.76	63293.74	5	NO.804	NO.8	1	148.1	80371.38
1	NO.153	NO.1	1	135.3	36541.22	5	NO.153	NO.1	1	105.4	58000.58	1	NO.805	NO.8	1	53.92	52303.42	5	NO.805	NO.8	1	91.42	51845.36
1	NO.154	NO.1	1	68.68	68404.94	5	NO.154	NO.1	1	108.56	89150.54	1	NO.808	NO.8	1	75.6	50655.5	5	NO.808	NO.8	1	82.24	58926.46
1	NO.155	NO.1	1	87.44	37439.7	5	NO.155	NO.1	1	87.92	58034.64	1	NO.809	NO.8	1	104.18	44708.08	5	NO.809	NO.8	1	103.42	61541.3
1	NO.156	NO.1	1	107.3	53483.88	5	NO.156	NO.1	1	127.52	71453.98	1	NO.811	NO.8	1	102.86	38044.42	5	NO.811	NO.8	1	102.92	42830.26
1	NO.157	NO.1	1	116.68	74873.16	5	NO.157	NO.1	1	123.14	104171.8	1	NO.812	NO.8	1	88.88	29293.78	5	NO.812	NO.8	1	66.82	41643.1
1	NO.158	NO.1	1	79.4	27644.54	5	NO.158	NO.1	1	47.36	34844.6	1	NO.813	NO.8	1	99.2	52320.64	5	NO.813	NO.8	1	97.7	50233.74
1	NO.162	NO.1	1	235.98	128556.78	5	NO.162	NO.1	1	231.16	132301.8	1	NO.814	NO.8	1	72.76	36298.54	5	NO.814	NO.8	1	71.02	48131.92
1	NO.164	NO.1	1	115.62	44833.46	5	NO.164	NO.1	1	99.56	44289.38	1	NO.816	NO.8	1	60.32	20551.1	5	NO.816	NO.8	1	60.38	20481.08
1	NO.165	NO.1	1	245.44	130819.18	5	NO.165	NO.1	1	169.2	115818.6	1	NO.817	NO.8	1	43.94	32386.92	5	NO.817	NO.8	1	75.16	28271.12
1	NO.166	NO.1	1	156.28	46018.38	5	NO.166	NO.1	1	126.56	49509.96	1	NO.820	NO.8	1	34.72	13302.84	5	NO.820	NO.8	1	29.46	11139.44

WK	StoreID	Division	Position	Hours	Sales
1	NO.168	NO.1	1	59.94	42647.58
1	NO.170	NO.1	1	275.04	165924.16
1	NO.171	NO.1	1	69.32	41674.34
1	NO.173	NO.1	1	38.2	34496.28
1	NO.174	NO.1	1	47.28	21550.34
1	NO.175	NO.1	1	174.12	80549.38
1	NO.176	NO.1	1	89.82	49312.38
1	NO.177	NO.1	1	230.44	91730.4
1	NO.179	NO.1	1	24	29319.2
1	NO.180	NO.1	1	41.58	37527.18
1	NO.181	NO.1	1	56.54	31145.76
1	NO.182	NO.1	1	254.02	108770.42
1	NO.183	NO.1	1	93.74	74727.86
1	NO.186	NO.1	1	200.94	33359.9
1	NO.187	NO.1	1	167.36	70376.78
1	NO.188	NO.1	1	236.7	96858.6
1	NO.194	NO.1	1	129.16	61547.82
1	NO.195	NO.1	1	73.84	30301.16
1	NO.200	NO.2	1	217.42	92838.8
1	NO.201	NO.2	1	119.08	76473.22
1	NO.202	NO.2	1	257.26	96648.5
1	NO.203	NO.2	1	182.44	68019.5
1	NO.204	NO.2	1	131	84719.76

WK	StoreID	Division	Position	WorkHours	Sales
5	NO.168	NO.1	1	80.54	58908.28
5	NO.170	NO.1	1	232.94	163253
5	NO.171	NO.1	1	35.18	53584.18
5	NO.173	NO.1	1	44.94	34301.48
5	NO.174	NO.1	1	75.06	21366.32
5	NO.175	NO.1	1	151.9	82024.68
5	NO.176	NO.1	1	51.14	49795.64
5	NO.177	NO.1	1	132.76	84422.64
5	NO.179	NO.1	1	34.3	28998.48
5	NO.180	NO.1	1	38.84	43911.36
5	NO.181	NO.1	1	56.5	33427.24
5	NO.182	NO.1	1	178.04	114663.3
5	NO.183	NO.1	1	176.76	78153
5	NO.186	NO.1	1	199.08	44944.4
5	NO.187	NO.1	1	317.58	115827.8
5	NO.188	NO.1	1	291.04	122728.5
5	NO.194	NO.1	1	252.14	67947.96
5	NO.195	NO.1	1	85.52	33845.64
5	NO.200	NO.2	1	249.62	101215
5	NO.201	NO.2	1	46.54	40632.16
5	NO.202	NO.2	1	170.48	111481.9
5	NO.203	NO.2	1	197.22	67240.02
5	NO.204	NO.2	1	217.2	109252.1

WK	StoreID	Division	Position	Hours	Sales
1	NO.822	NO.8	1	66.02	31558.2
1	NO.824	NO.8	1	92.18	44975.5
1	NO.828	NO.8	1	73.86	17499.96
1	NO.830	NO.8	1	33.54	20302.32
1	NO.832	NO.8	1	38.02	30276.74
1	NO.835	NO.8	1	55.98	25023.24
1	NO.837	NO.8	1	36.88	17174
1	NO.842	NO.8	1	42.14	32905.14
1	NO.847	NO.8	1	35.48	18565.64
1	NO.848	NO.8	1	86.72	15198.7
1	NO.905	NO.9	1	30.48	28565.44
1	NO.906	NO.9	1	35.46	34670.2
1	NO.919	NO.9	1	49.46	37947.4
1	NO.920	NO.9	1	72.46	47694.36
1	NO.922	NO.9	1	38.94	48752.72
1	NO.930	NO.9	1	85.22	35796.38
1	NO.938	NO.9	1	37.4	32978.84
1	NO.946	NO.9	1	38.92	28291.66
1	NO.947	NO.9	1	84.54	19159.6
1	NO.953	NO.9	1	32.64	30692.24
1	NO.958	NO.9	1	93.48	48923.26
1	NO.965	NO.9	1	35.42	40124.56
1	NO.968	NO.9	1	80.54	68217.84
1	NO.976	NO.9	1	33.38	27223.96

WK	StoreID	Division	Position	Hours	Sales
5	NO.822	NO.8	1	76.64	32904.18
5	NO.824	NO.8	1	107.26	50382.54
5	NO.828	NO.8	1	90.32	14660.46
5	NO.830	NO.8	1	32.3	26366.22
5	NO.832	NO.8	1	37.74	39462.36
5	NO.835	NO.8	1	75.76	25645.3
5	NO.837	NO.8	1	59.64	24559.7
5	NO.842	NO.8	1	83.1	47270.28
5	NO.847	NO.8	1	36.94	21877.06
5	NO.848	NO.8	1	117.98	14046.16
5	NO.905	NO.9	1	30.06	23313.04
5	NO.906	NO.9	1	30.96	31702.68
5	NO.919	NO.9	1	47.72	41517.92
5	NO.920	NO.9	1	88.26	41563.08
5	NO.922	NO.9	1	37.42	39341.64
5	NO.930	NO.9	1	70.36	28368.84
5	NO.938	NO.9	1	41.44	22934.66
5	NO.946	NO.9	1	48.6	26294.4
5	NO.947	NO.9	1	48.66	21930.34
5	NO.953	NO.9	1	32.3	22080.38
5	NO.958	NO.9	1	56.7	33414.78
5	NO.965	NO.9	1	47.38	24490.44
5	NO.968	NO.9	1	89.78	55418.74
5	NO.976	NO.9	1	33.38	17069.62

Bibliography

[1] Andridge R.R. and Little R.J.A. (2010) A review of hot deck imputation for survey non-response *International Statistical Review* **78**,1, 40-64.

[2] Chaubey Y.P. and Sen P.K. (2009) On the selection of the smoothing parameter in Poisson smoothing of histogram estimator: computational aspects *Pak. J. Statist.* **25**,4, 385-401.

[3] Chaubey Y.P., Laïb N. and Sen A. (2010) Generalised kernel smoothing for non-negative stationary ergodic processes *Journal of Nonparametric Statistics* **22**,8, 973-97.

[4] Chen J., Rao J.N.K. and Sitter R.R. (2000) Efficient random imputation for missing data in complex surveys *Statistica Sinica* **10**, 1153-69.

[5] Cheng P.E. and Wei L.J. (1986) Nonparametric inference under ignorable missing data process and treatment assignment *International Statistical Symposium*, Taipei, **1**, 97-112.

[6] Cochran W.G. (1977) *Sampling Techniques* John Wiley & Sons.

[7] Dang X. and Serfling R. (2009) A numerical study of multiple imputation methods using nonparametric multivariate outlier identiers and depth-based performance criteria with clinical laboratory data *Journal of Statistical Computation and Simulation* **81**, 5, 547-60.

[8] Efromovich S. (1999) *Nonparametric Curve Estimation. Methods, Theory, and Applications* Springer.

[9] Efromovich S. (2012) Nonparametric regression with missing data: theory and applications *Department of Mathematical Sciences, UTDallas, TX 75080, USA*

[10] Feller W. (1968) *An Introduction to Probability Theory and Its Application Volume I* (Third Edition) John Wiley & Sons.

[11] Feller W. (1971) *An Introduction to Probability Theory and Its Application Volume II* (Second Edition) John Wiley & Sons.

[12] Härdle W. (1990) *Applied nonparametric regression* Cambridge University Press.

[13] Horton N. and Lipsitz S.R. (2001) Multiple Imputation in Practice: Comparison of Software Packages for Re-

gression Models With Missing Variables *The American Statistician* **55**, 3, 244-54.

[14] IBM (2011) *IBM SPSS Missing Values 20* IBM Corporation.

[15] Kim J.K., Brick J.M., Fuller W.A. and Kalton G. (2006) On the bias of the multiple-imputation variance estimator in survey sampling *J.R. Statist. Soc* B **68**,3, 509-21.

[16] Krishnamoorthy K., Mallick A. and Mathew M. (2009) Model-based imputation approach for data analysis in the presence of non-detects *Ann. Occup. Hyg.* **53**,3, 249-63.

[17] Little R.J.A. and Rubin D.B. (2002) *Statistical Analysis with Missing Data* 2nd ed. New York: Wiley.

[18] Lohr S.L. (2010) *Sampling: Design and Analysis (Second Edition)* Boston MA: Brooks/Cole.

[19] Luengo J., Garcia S. and Herrera F. (2012) On the choice of the best imputation methods for missing values considering three groups of classification methods *Knowl. Inf. Syst.* **32** 77-108

[20] Ning J. and Cheng P.E. (2012) A comparison study of nonparametric imputation methods *Statistics and Computing* **22**, 1, 273-85

[21] Pan W. and Connett J.E. (2001) A Multiple Imputation Approach to Linear Regression with Clustered Censored Data *Lifetime Data Anal.* **7** 77-108

[22] Quintano C., Castellano R. and Rocca A. (2010) Influence of Outliers on Some Multiple Imputation Methods *Metodološki zvezki* **7**, 2, 111-23

[23] Raghunathan T. E., Lepkowski J.M., Van Hoewyk J. and Solenberger P. (2001) A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models *Survey Methodology* **27**, 1, 85-95

[24] Rao J.N.K. and Sitter R.R. (1995) Variance estimation under two-phase sampling with application to imputation for missing data *Biometrika* **82**,2, 453-60.

[25] Rice J.A. (2007) *Mathematical Statistics and Data Analysis* `(Third Edition)` Thomson Brooks/Cole.

[26] Rubin D.B. (2004) *Multiple Imputation for Nonresponse in Surveys* Hoboken NJ: Wiley-Interscience.

[27] Sande I.G. (1982) Imputation in surveys: coping from reality *The American Statistician* **36**,3, Part 1, 145-52.

[28] Schafer J.L. (2000) *Analysis of Incomplete Multivariate Data* Chapman & Hall/CRC.

[29] Singh S. and Deo B. (2003) Imputation by power transformation *Statist. Papers* 555-79.

[30] Singh S. and Valdes S.R. (2009) Optimal method of imputation in survey sampling *Applied Mathematical Sciences* **3**,35, 1727-37.

[31] Stekhoven D.J. and Bühlmann P. (2012) MissForest - nonparametric missing value imputation for mixed-type data *Bioinformatics/Oxford* **28**, 1, 112-8

[32] Wang D. and Chen S.X. (2006) Nonparametric imputation of missing values for estimating equation based inference *CiteSeerX/IaState*

[33] Wayman Y.C. (2000) Multiple Imputation for Missing Data: Concepts and New Development *Annual Meeting of the American Educational Research Association, Chicago, IL.*

[34] Yuan J.C. (2003) Multiple Imputation For Missing Data: What Is It And How Can I Use It? *Proceeding of the Twenty-Fifth Annual SAS®Users Group International Conference* Paper **267**.