ON THE INTERPOLATION OF MISSING DEPENDENT VARIABLE OBSERVATIONS

A Thesis
Presented to
The Graduate Faculty of The University of Akron

In Partial Fulfillment
of the Requirements for the Degree
Master of Economics

Alexander Mark Medvedeff
May, 2008
ON THE INTERPOLATION OF MISSING DEPENDENT VARIABLE OBSERVATIONS

Alexander Mark Medvedeff

Thesis

Approved

Advisor
Dr. Steven Myers

Co-Advisor
Dr. Gasper Garofalo

Department Chair
Dr. Mike Nelson

Accepted:

Dean of the College
Dr. Ronald Levant

Dean of the Graduate School
Dr. George Newkombe

Date
ABSTRACT

This paper utilizes a Monte Carlo experiment to simulate random missing observations in a dataset in order to analyze the effects of how various techniques to compensate for missing dependent variable observations behave in time series regression analysis. Reduced Sample, Modified Zero Order, and First Order techniques are tested with authentic data and simulated population datasets. The size of the datasets and the relative percentage of observations simulated as missing are changed in order to investigate the sensitivity of results to different dataset conditions. Each combination of data-type, size of dataset, percentage of missing observations, and model specification are regressed 1,000 times. Results are compared to a control or “full information” regression in order to analyze the bias and efficiency of each method tested.

Results indicate that the Reduced Sample method should be the preferred solution for dealing with missing dependent variables in time series regression, as it consistently produces the least biased and most efficient estimates. The results also indicate a small degree of sensitivity to model specification, size of dataset and the percentage of observation simulated as missing.
ACKNOWLEDGEMENTS

I would like to thank my family. To my parents, without your continual support throughout my life and academic career I could not be where I am today. You have instilled in me a love of learning that will last for the rest of my life. To my sister Megan, you have always set a high standard for me to live up to, and for that I am very thankful. Without it, I would not have had the motivation to always work harder and strive for nothing less than perfection (or at least outperforming you!).

I would also like to thank my friends, both old and new who have helped me along the way. You have all played an enormous role in my academic, intellectual, and personal development and without you I would not be the person that I am today. To those of you who were there, I miss the late night conversations and arguments about everything and nothing. Despite the blistering heat and freezing cold of poor off-campus living, I will always remember those as some of the best times of my life.

Finally, I would like to thank Dr. Steven Myers for the important role that he played in the development of this project from start to finish, and for encouraging me to present at the Midwest Economics Association conferences.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
</tbody>
</table>

CHAPTER

I. INTRODUCTION ................................................................. 1

II. LITERATURE REVIEW ......................................................... 4

  Reduced Sample Regression .................................................. 5
  Zero Order Regression ......................................................... 6
  Modified Zero Order Regression ........................................... 7
  First Order Regression ....................................................... 8
  Concerns Regarding Monte Carlo Analysis ................................ 9
  Discussion of Missing Observations and Interpolation Methodologies ........................................ 10

III. DATA AND METHODOLOGY ............................................. 13

  Methodology ................................................................. 13
  Judgment Criterion for Results ........................................... 15
  Data .................................................................................. 16

IV. RESULTS AND DISCUSSION ........................................... 19

  Results for Authentic Data .................................................. 19
  Results for Simulated Data .................................................. 27
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Control Regression Results: Authentic Data</td>
<td>20</td>
</tr>
<tr>
<td>4.2</td>
<td>Summary Statistics for Frequency Distributions Based on Authentic Data, OLS Specification</td>
<td>20</td>
</tr>
<tr>
<td>4.3</td>
<td>Summary Statistics for Frequency Distributions Based on Authentic Data, AR2 Specification</td>
<td>21</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary of Efficiency: Percent of Estimates within ( + / - ) One-Half the Standard Error from Control Estimate, Authentic Data, OLS Specification</td>
<td>26</td>
</tr>
<tr>
<td>4.6</td>
<td>Control Regression Results: Simulated Data</td>
<td>28</td>
</tr>
<tr>
<td>4.7</td>
<td>Summary Statistics for Frequency Distributions Based on Simulated Data, OLS Specification</td>
<td>29</td>
</tr>
<tr>
<td>4.8</td>
<td>Summary Statistics for Frequency Distributions Based on Authentic Data, AR2 Specification</td>
<td>29</td>
</tr>
<tr>
<td>4.9</td>
<td>Summary of Efficiency: Percent of Estimates within ( + / - ) One-Half the Standard Error from Control Estimate, Simulated Data, OLS Specification</td>
<td>33</td>
</tr>
<tr>
<td>4.10</td>
<td>Summary of Efficiency: Percent of Estimates within ( + / - ) One-Half the Standard Error from Control Estimate, Simulated Data, AR2 Specification</td>
<td>34</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Frequency Distribution for Reduced Sample Method: Authentic Data, 100 observations, 5% Simulated as Missing, OLS Specification</td>
<td>23</td>
</tr>
<tr>
<td>4.2</td>
<td>Frequency Distribution for First Order Method: Authentic Data, 100 observations, 5% Simulated as Missing, OLS Specification</td>
<td>23</td>
</tr>
<tr>
<td>4.3</td>
<td>Frequency Distribution for Reduced Sample Method: Authentic Data, 200 Observations, 10% Simulated as Missing, AR2 Specification</td>
<td>24</td>
</tr>
<tr>
<td>4.4</td>
<td>Frequency Distribution for First Order Method: Authentic Data, 200 Observations, 10% Simulated as Missing, AR2 Specification</td>
<td>25</td>
</tr>
<tr>
<td>4.5</td>
<td>Frequency Distribution for Reduced Sample Method: Simulated Data, 100 observations, 5% Simulated as Missing, OLS Specification</td>
<td>30</td>
</tr>
<tr>
<td>4.6</td>
<td>Frequency Distribution for First Order Method: Simulated Data, 100 observations, 5% Simulated as Missing, OLS Specification</td>
<td>31</td>
</tr>
<tr>
<td>4.7</td>
<td>Frequency Distribution for Reduced Sample Method: Simulated Data, 200 Observations, 10% Simulated as Missing, AR2 Specification</td>
<td>32</td>
</tr>
<tr>
<td>4.8</td>
<td>Frequency Distribution for First Order Method: Simulated Data, 200 Observations, 10% Simulated as Missing, AR2 Specification</td>
<td>32</td>
</tr>
<tr>
<td>A.1</td>
<td>Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 5% Simulated as Missing, OLS Specification</td>
<td>42</td>
</tr>
<tr>
<td>A.2</td>
<td>Frequency Distribution for Reduced Sample Method: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification</td>
<td>43</td>
</tr>
<tr>
<td>A.3</td>
<td>Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification</td>
<td>43</td>
</tr>
<tr>
<td>A.4</td>
<td>Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification</td>
<td>44</td>
</tr>
</tbody>
</table>
A.5  Frequency Distribution for Reduced Sample Method: Authentic Data,
     200 Observations, 5% Simulated as Missing, OLS Specification.................44

A.6  Frequency Distribution for Modified Zero Order Method: Authentic Data,
     200 Observations, 5% Simulated as Missing, OLS Specification.................45

A.7  Frequency Distribution for First Order Method: Authentic Data,
     200 Observations, 5% Simulated as Missing, OLS Specification.................45

A.8  Frequency Distribution for Reduced Sample Method: Authentic Data,
     200 Observations, 10% Simulated as Missing, OLS Specification.................46

A.9  Frequency Distribution for Modified Zero Order Method: Authentic Data,
     200 Observations, 10% Simulated as Missing, OLS Specification.................46

A.10 Frequency Distribution for First Order Method: Authentic Data,
     200 Observations, 5% Simulated as Missing, OLS Specification.................47

B.1  Frequency Distribution for Reduced Sample Method: Authentic Data,
     100 Observations, 5% Simulated as Missing, AR2 Specification.................48

B.2  Frequency Distribution for Modified Zero Order Method: Authentic Data,
     100 Observations, 5% Simulated as Missing, AR2 Specification.................49

B.3  Frequency Distribution for First Order Method: Authentic Data,
     100 Observations, 5% Simulated as Missing, AR2 Specification.................49

B.4  Frequency Distribution for Reduced Sample Method: Authentic Data,
     100 Observations, 10% Simulated as Missing, AR2 Specification.................50

B.5  Frequency Distribution for Modified Zero Order Method: Authentic Data,
     100 Observations, 10% Simulated as Missing, AR2 Specification.................50

B.6  Frequency Distribution for First Order Method: Authentic Data,
     100 Observations, 10% Simulated as Missing, AR2 Specification.................51

B.7  Frequency Distribution for Reduced Sample Method: Authentic Data,
     200 Observations, 5% Simulated as Missing, AR2 Specification.................51

B.8  Frequency Distribution for Modified Zero Order Method: Authentic Data,
     200 Observations, 5% Simulated as Missing, AR2 Specification.................52

B.9  Frequency Distribution for First Order Method: Authentic Data,
     200 Observations, 5% Simulated as Missing, AR2 Specification.................52
B.10 Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 10% Simulated as Missing, AR2 Specification..........................53

C.1 Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 5% Simulated as Missing, OLS Specification..............................54

C.2 Frequency Distribution for Reduced Sample Method: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification..............................55

C.3 Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification..............................55

C.4 Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification..............................56

C.5 Frequency Distribution for Reduced Sample Method: Authentic Data, 200 Observations, 5% Simulated as Missing, OLS Specification..............................56

C.6 Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, OLS Specification..............................57

C.7 Frequency Distribution for First Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, OLS Specification..............................57

C.8 Frequency Distribution for Reduced Sample Method: Authentic Data, 200 Observations, 10% Simulated as Missing, OLS Specification..............................58

C.9 Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 10% Simulated as Missing, OLS Specification..............................58

C.10 Frequency Distribution for First Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, OLS Specification..............................59

D.1 Frequency Distribution for Reduced Sample Method: Authentic Data, 100 Observations, 5% Simulated as Missing, AR2 Specification..............................60

D.2 Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 5% Simulated as Missing, AR2 Specification..............................61

D.3 Frequency Distribution for First Order Method: Authentic Data, 100 Observations, 5% Simulated as Missing, AR2 Specification..............................61

D.4 Frequency Distribution for Reduced Sample Method: Authentic Data, 100 Observations, 10% Simulated as Missing, AR2 Specification..............................62
D.5 Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, AR2 Specification..................62

D.6 Frequency Distribution for First Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, AR2 Specification......................63

D.7 Frequency Distribution for Reduced Sample Method: Authentic Data, 200 Observations, 5% Simulated as Missing, AR2 Specification.......................63

D.8 Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, AR2 Specification.......................64

D.9 Frequency Distribution for First Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, AR2 Specification.......................64

D.10 Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 10% Simulated as Missing, AR2 Specification....................65
CHAPTER I
INTRODUCTION

Any research that relies on empirical investigations based in real world data is vulnerable to problems arising from imperfect datasets. One such issue is missing observations; a topic which received attention in the 1950’s and 1960’s but has since been mostly overlooked. Much of this research involved testing a single technique to compensate for missing data within a single dataset, but unfortunately the power of these investigations was limited by the efficiency of technology available.

Since that time computers have become substantially more efficient and the knowledge bank of statistical relationships has increased dramatically. However, these original ideas and methods have not been re-examined despite these computational advances and consequently not much is known about the characteristics of suggested interpolation techniques. The goal of this paper is two fold: to empirically test compensatory techniques for missing data in a dependent variable framework for time series analysis, and to present the results in a format more accessible to researchers.

Having an incomplete array of variable observations can occur for a variety of reasons, including but not limited to irregular or sporadic collection, varied frequencies of collection, non-responses in surveys, proprietary data, poor data collections standards and incomplete data sources. Regardless of the cause, the concept of randomness in the
missing data must be carefully considered\textsuperscript{i}. A researcher needs to differentiate between malicious and benign causal factors for missing observations because these cases must be treated differently. If there was suspicion that values were intentionally left out (malicious), any attempt to interpolate or recreate the missing values might introduce bias or reduce efficiency by not controlling for the underlying cause. In such cases research should not continue. However, if the cause of the missing variable observations appears to be benign, research may proceed with caution.

A second characteristic of randomness suggests that missing values should be scattered throughout the dataset without any recognizable pattern. If there seems to be a pattern, then interpolation may again introduce bias or inefficiency. The techniques tested in this paper are not intended to correct for such a predictable structure of missing observations. Approaches to datasets with predictably missing observations will be discussed in greater detail in Chapter II.

Assuming a dataset the meets acceptable criteria for randomness, applying standard statistical practices to datasets with missing dependent variable observations overlooks parameter stability. The effects of interpolation on the overall quality of the model remain unclear without a basis for comparison, which a single sample cannot provide. By using a repeated sampling Monte Carlo technique with a focus on time series analysis\textsuperscript{ii}, this paper investigates the effects of three different interpolation techniques.

\textsuperscript{i} Granted, it is the case that \textit{truly} random missing observations do not exist because there is always an underlying reason (though it is not always known), but this misses the point.

\textsuperscript{ii} When missing observations occur in cross sectional analysis, standard practice is to simply delete the incomplete observations from the dataset. It is likely impossible that any other method will outperform this standard reduced sample OLS method and as such, cross sectional models will not be tested. For more details, see the Reduced Sample Regression section in Chapter II.
Utilizing a simple Keynesian consumption function with authentic data, and simulated population datasets, this paper will test the sensitivity of results to the size of the dataset and the relative percentage of missing data. Each set of results will then be compared to a control “full information” regression to identify inherent bias and inefficiency.

This paper investigates the effects of interpolating dependent variable observations by pursuing the contexts in which interpolation is appropriate and the relative performance of three different techniques. It is hoped that by developing new knowledge and a broader understanding of these applications the results of this paper will make the interpolation of dependent variable observations more reliable and accessible.
CHAPTER II
LITERATURE REVIEW

Several interpolation techniques have been discussed in the literature for dealing with missing variable observations: Reduced Sample OLS, Zero Order Regression, Modified Zero Order Regression and First Order Regression\textsuperscript{iii}. In the literature, these methods are presented in a general context, and no real differentiation is made between their application to time series and cross sectional research\textsuperscript{iv}.

Before turning to the details of these methods, consider a dataset of arbitrary size that has missing observations scattered throughout, and assume three subsets of this complete dataset: Group A is comprised of all complete observations, Group B contains observations with only the dependent variable observation missing, and Group C contains observations with only a missing independent variable observation. This decomposition of a complete dataset is utilized by Maddala (1977) and Greene (1993) for simplicity and will be used throughout this section to help illustrate each technique.

\textsuperscript{iii} Various Maximum Likelihood estimation techniques have also been suggested. Buck (1960) warns that these methods are impractical for anything outside of two variable models. Furthermore, much of this research is focused on missing independent variables. Subsequently, these techniques are omitted from this paper, as the goal is to understand the use of interpolation for dependent variables on a practical level. The interested reader should see the references section of this paper for some of the original publications.

\textsuperscript{iv} This paper will provide more recent citations for these methods because the text-book literature where they are presented is far more accessible (and readable) than the original articles. However, references to many of the original articles are contained in the bibliography of this paper.
While the techniques presented in this section are applicable to both missing dependent and independent variable observations, this investigation is concerned with the interpolation of dependent variable observations only. It is therefore unnecessary to consider Group C, which is comprised of observations with missing independent variable values, and hence it will be ignored leaving the focus on Groups A and B.

*Reduced Sample Regression (RSOLS)*

Reduced sample regression involves simply deleting all rows with missing observations and running regressions on the remaining ones, which comprise all of Group A. Maddala (1977) describes this as a poor method, as it generally undesirable to simply discard observations for which an observation is missing. This is especially true in situations when data must be purchased, in which case researchers should seek to extract possible information from the dataset.

This method may be questionable in its application to time series analysis, as it removes years from the analysis in a non-uniform way. When observations are simply removed, the dataset is collapsed to fill the resulting holes. This reorganization of the dataset risks leaving large jumps in the data. To illustrate this, consider annual data from 1960-2000 for any arbitrary time series model. If for example, the years 1980 and 1981 are missing for the dependent variable, this method suggests deleting the complete observation, which reorganizes the dataset so that the data for 1979 is followed directly by 1982. If there is some obvious trend in the data, placing 1979 directly next to 1982

---

The author considered presenting a “Time Series Adjusted Reduced Sample Method” which would have involved coding *all* values in a row as missing whenever the dependent variable is missing. However, this method would result in unreliable estimates when correcting for autocorrelation, because the lagged error structure would be based on a blank observation.
makes it likely that there will be an abnormal jump from one year to the next, potentially biasing the results.

In contrast to time series data where the ordering of the dataset clearly matters, data for cross sectional analysis is independent and identically distributed (iid), implying that any shuffling or reorganization of the rows of data will not bias the results. As a result, reorganizing the dataset by implementing this Reduced Sample method will not bias the results of the regression. Because this method is considered standard practice for cross sectional analysis, the only instance when it would not be used would be when having a small number of observations increases the likelihood of inducing small sample bias. However, since the other interpolation techniques (yet to be discussed) rely on the observed data, the interpolated values would be biased, and hence final regression results would still be biased. This same argument can be applied for time series regressions, indicating that whenever the number of complete observations is insufficient to avoid the small sample bias, research should not proceed.

Zero Order Regression (ZO)

Zero Order regression involves calculating the mean of the observed dependent variables from Group A and inserting that value into each missing observation from Group B (Greene 1993). At this point, having filled in all missing observations, standard regression methods can be implemented on a dataset comprised of complete observations, which is the union of Groups A and B. Greene suggests that this method is statistically identical to dropping the incomplete data, though yields a lower $R^2$.

Greene (1993) goes on to state that any method that claims to produce unbiased estimates by filling in missing values of the dependent variable must do it in a way so
that a regression using only filled observations would still produce an unbiased estimator. With this in mind he proves the Zero Order method as described fails this test. This proof will not be presented here, and the interested reader should seek it out in its original context.

Additionally Greene offers the following, more intuitive, explanation for the Zero Order method failing to produce an unbiased estimator. (Paraphrasing) Consider a two variable model in which X is the independent variable and Y is the dependent variable. Suppose then that the estimated regression coefficient, β, is positive, in which case large values of X are associated with large values of Y. However, if all missing Y observations are replaced with the mean obtained from Group A, this association of large X with large Y is lost (for large X). Hence, the independent variable and the error term for these observations are not independent, and the OLS estimator for the slope and intercept will be biased. In light of this known bias, the Zero Order method will not be tested, and should not be considered as a viable technique to compensate for missing variable observationsvi.

*Modified Zero Order Regression (MZO)*

Instead of using the mean, the Modified Zero-Order regression discussed by Maddala (1977) and Greene (2003) involves filling all missing dependent variable observations (Group B) with zero, and creating a dummy variable that takes the value of -1 for all observed dependent variable observations, and 0 for all missing values. Having modified the dataset to be complete (with the addition of a new variable), standard

---

vi This method was included in preliminary analysis for the sake of demonstrating this known bias. The results indicated a consistent depression of the magnitude of both slope and intercept terms. However, in light of its *known* bias, there is no reason to include the analysis in this document.
regression techniques can be applied with the inclusion of the dummy as an independent variable. This new dummy variable alters the specification of the model, and under the assumption of proper model specification the OLS estimator will be biased. However, proper model specification may not always be a reasonable assumption, so the Modified Zero Order method may be valid in certain circumstances.

This method is statistically equivalent to substituting the mean of the dependent variable, like in the standard Zero-Order regression. Unlike the Zero Order Method however, no proof has been offered to demonstrate that the Modified Zero Order Method is inherently biased, and hence this method will be tested in this paper.

First Order Regression (FO)

First-Order regression involves running an Ordinary Least Squares regression on subset A of the original dataset and extracting the parameter estimates (Kelejian 1969) (Maddala 1977)\(^\text{vii}\). These estimates are then multiplied by the corresponding independent variables in Group B and summed in order to estimate a value for the missing independent variable. Standard regression techniques can then be preformed on the modified dataset. However, such regressions on the modified dataset must be based on the observed values from Group A and the estimated observations from Group B. If all the observations used were estimated, then a direct linear relationship would be established and such a regression would fail.

\(^\text{vii}\) A Modified First Order Regression is suggested by Maddala, which involves estimating all missing values of X based on observed values of Y, and simply discarding all observations with missing Y values. However, this Modified First Order Regression simply breaks down into the Reduced Sample Regression method, given this papers assumption that all values of X are observed (i.e. Group C is empty, or at least ignored).
Greene (2003) warns that the First-Order method should be used with caution, as it is contingent on having proper model specification in the original OLS model from which the interpolated values will be based. Under the assumption of a properly specified model, the interpolated values will be built out of unbiased parameter estimates, and will be pass the test for being unbiased (see the Zero Order Regression discussion above). Greene claims that while this method seems to bring a gain in efficiency, it may be illusory because it is also then necessary to account for additional variation obtained through the estimated values of Y.

*Concerns Regarding Monte Carlo Analysis*

Greene (2003) warns that little is known about the previously described methods in a dependent variable context. What little is known has been derived from Monte Carlo experiments, the results of which are generally difficult to generalize, as they are based on fixed patterns of missing data, or specific datasets.

Haitovsky (1968) tested the Reduced Sample and Zero Order Regression methods for eight different models and two fixed patterns of missing data, repeating each combination ten times. However, Haitovsky did not standardize the number of missing observations in the simulation, which is an aspect of the Monte Carlo method that this paper improves upon. His results indicated that the Reduced Sample method outperforms the Zero Order method.

Haitovsky’s research seems to have been limited by the capabilities of computer technology at the time, as a Monte Carlo experiment based on 10 repetitions is dangerously inadequate. In light of this, Green’s concerns seem to be well warranted.
Fortunately since the late 1960’s computers have advanced quite a bit, and the limitations that existed for Haitovsky no longer apply. Though it will be discussed in greater detail later in the methodology, this research seeks to avoid these concerns by increasing the number of repetitions from 10 to 1,000 and by not applying a fixed pattern for deletion. In doing so the results obtained from this investigation will be more easily generalized, as they will not suffer from these weaknesses.

Mariano et al (2000) warns that care needs to be taken when dealing with Monte Carlo analysis as it is difficult to generalize the results to cases that have not been simulated. The methodology developed for this investigation is intended to avoid this problem by drawing conclusions on a wide variety of different simulations.

Hendry (1993) warns that estimates obtained via Monte Carlo methods generally lack precision, even when based on a large number of samples. For this investigation all of the models tested with interpolation techniques will be compared to a control or “full information” model in order to illustrate how consistent each method is. Furthermore, it is exactly this degree of precision / imprecision of parameter estimates that concerns this investigation. The results obtained will be used to understand the performance of various interpolation techniques in a statistical sense and not for generalization or interpretation in an economic sense.

Discussion of Missing Observations and Interpolation Methodologies

This paper will now turn its focus to discuss several specific causes of missing variable observations and why research should or should not proceed for each. First, consider an incomplete dataset resulting from sporadic or irregular data collection. Because not all nations of the world have data collection regulations as strict as those in
the United States, this type of infrequency or irregular collection is often a problem when dealing with international research and developing countries. Consider a hypothetical case study of poverty for a country such as Bolivia, which is still a developing nation and lacks the degree of political infrastructure of the United States. It may be the case the Bolivian government collects poverty data in an on-again off-again manner, for example, the data may only exist for the years 1985, 1989, 1990, 1993, 1998, 2000, 2001, 2004, and 2006. In this situation there is no recognizable pattern to the observed years, and any attempt for interpolation would not control for the underlying cause for this sporadic collection frequency, thereby running the risk of introducing bias. However, if the data was observed for the years 1985, 1989, 1993, 1997, 2001 and 2005 (every four years), there is a clear pattern to this collection method and in this circumstance the problem with empirical work may simply be an aggregation / disaggregation issue, in which case the researcher should reference the existing body of literature surrounding methods for assorted time aggregations within a single dataset.

Another common cause of missing data observations comes from non-responses in surveys. Such a situation needs to be examined carefully, as one of two possibilities is likely. If there seems to be no pattern to the missing responses then it may be possible to consider the missing observations as random and research may proceed with caution. However, if there is some systematic pattern that exists in who failed to respond to

---

viii A major motivation for this project was a similar case study involving poverty in Mexico. For a real-life example of sporadic data points, the interested reader should seek poverty statistics for Mexico in the Statistical Abstract of Latin America, World Development Indicators Online, and Penn World Table.

ix It may be possible to apply this argument to the use of the reduced sample OLS technique with Time series, assuming that the resulting dataset had a fixed pattern for missing observations.
questions then it is possible to continue, although alternative methods are likely necessary. This type of situation arises out of a poorly structured survey, questions that were only asked to a certain subgroup of respondents, or from questions that are perceived as too personal or offensive to particular groups (Greene 2003). In these cases, a Two-Stage Least Squares regression method similar to the work of Heckman (1979) may be appropriate, where the first stage investigates the probability of responding, from which the Inverse Mills Ratio is extracted and included in the second stage as an independent variable\(^x\).

Lastly, consider missing observations that arise from poor data sources. In such situations research should not proceed until all other possible sources have been exhausted, and even then research should only proceed with interpolation if there is no cause to suspect that the missing observations are non-random, as discussed in the introduction.

Empirical work based on interpolated dependent variable observations can only take place in truly random circumstances, or when there is a clear structural break in the data, such as from surveys. Knowing this, the methods used in this investigation have been designed to simulate as best as possible situations of true randomness, thereby replicating conditions under which interpolation is permissible.

\(^x\) This proposed method is speculation; as such situations are not tested in this investigation. This investigation focuses on time series analysis and not on cross sectional analysis, so testing this method is left for future research.
CHAPTER III

DATA AND METHODOLOGY

Methodology

Consider a dataset in the form of an $n$ by $k$ matrix, where $n$ denotes the number of observations (rows) and $k$ the number of variables (columns), both dependent and independent. For the purposes of this section, this will be referred to as the original dataset. Before proceeding, this original dataset is used to run a standard OLS regression model which is properly specified. The results from this regression will serve as a control, or a baseline comparison for what is to come.

The next step is to generate a new dataset with a fixed quantity of random numbers, $x_i$ which will later be used to simulate missing observations. To be more precise, consider a column vector with $x$ generated observations, such that $x < n$, where $n$ refers to the number of observations in the original dataset. A uniform distribution is used for this number generating procedure because it creates an equal probability of any individual number being drawn. Any other type of distribution would generate predictable characteristics in the data thereby removing the random nature of what will become the missing observations.

---

$x_i$ The numbers generated via the UNIFORM procedure in SAS. Like any other computer based procedure, the numbers generated here are in fact pseudo-random. However, this paper will drop the qualification of pseudo and simply refer to them as random.
Using the congruential generator which utilizes a starting input seed value, random numbers are generated on the interval $[0, 1]$ (Fan et al 2002). All of the values of $x$ are then scaled upwards, mapping the random numbers drawn on the interval $[0,1]$ into the interval $[1,n]$. This transformation does not disturb the random nature of these values because each number is scaled in an identical manner and hence the transformation is uniform across all values. At this point all of the transformed values in the vector are rounded to integers so that they can be matched with an observation number in the original $n$ by $k$ matrix. This procedure will be repeated until there are 1,000 datasets comprised of computer generated random numbers.

Unfortunately it is possible that multiple values within the same dataset will be rounded to the same integer. In order to hold constant the percentage of missing observations, all datasets must have an identical number of unique randomly generated observations, making the repetition of random values an undesirable outcome within any given dataset. In order to assure the use of datasets without repetition, the previously described method will be repeated in excess of 1,000 times though only the first 1,000 datasets meeting this criterion will be selected for use.

Up to this point, this method has created 1,000 datasets containing exactly $x$ randomly generated numbers from a uniform distribution. Each of these datasets is then merged individually with the original $n$ by $k$ dataset forming a new $n$ by $k+1$ datasets, such that the value of each randomly generated integer corresponds to an observation number. For example, if the numbers 3, 9, and 24 were generated, then the 3 would be placed in the 3$^{rd}$ row, the 9 in the 9$^{th}$ row and the 24 in the 24$^{th}$ row, leaving all other spots empty. This process tags the dependent variable for deletion. Again, assuming that
the numbers 3, 9, and 24 are generated, the end result of this procedure is that in this particular unique dataset, the 3\textsuperscript{rd}, 9\textsuperscript{th}, and 24\textsuperscript{th} observations of the dependent variable will be deleted. The idea is that, when 3 random numbers are generated, the subsequent dataset will have exactly 3 random missing dependent observations, when 10 numbers are generated then the resulting dataset will have 10 random missing dependent variable observations etc. The end result is a group of datasets with a complete array of independent variable observations and a consistent number of missing dependent variable observations. Satisfying the criterion of randomness, these datasets are ready to have interpolation techniques applied.

*Judgment Criterion for Results*

Because it is difficult to directly compare model fit measurements such as the $R^2$ or Adjusted-$R^2$ for 1,000 regressions in a practical and meaningful way, additional judgment criterions need to be developed in order to measure the relative performance of each technique. To accomplish this, a frequency distribution will be constructed for the values of each dependent variable’s parameters, in which values are dropped into intervals of a fixed size. Under the OLS assumptions of homoscedasticity and proper model specification, it reasonable to assume a somewhat normal looking distribution for the parameter values developed through this method. Given constant variance, changes that come from interpolating values of the dependent variable should result in deviations both above and below the control, pending which observations are randomly selected as missing. Each of these distributions will be compared to the parameter obtained from the control regressions that will be run for each model on the full sample, by overlaying a
normal distribution created from the control regression’s parameter estimates and standard errors.

Haitovsky (1968) suggests the use of the Mean Square Error statistics from each regression as a judgment criterion. As such a method will reflect on both the efficiency and bias of a technique, summary statistics will be presented for the number observations that fall within (±/−) one-half of the standard error from the control regression estimate.

Data

This paper will utilize a simple Keynesian consumption function in which consumption is a function of same-period personal disposable income. This model has been chosen as the basis for empirical work, so as to make the assumption of proper model specification reasonable. Data comes from the FRED database managed by the Federal Reserve Bank of St. Louis and ranges from 1956-2005 on a quarterly basis and has been seasonally adjusted (200 observations). Additionally, a subset of this complete dataset will be used, which will range from 1981-2005 (100 observations). By using these datasets of different sizes, it is possible to examine how sensitive results based on interpolated data are to the overall size of the dataset. Additionally, in order to examine how sensitive the results based on interpolated data are to the relative percentage of missing data, these two datasets will be used, in accordance with the previously discussed method, to simulate conditions in which 5% and 10% of the data are missing. Each of these datasets will be used to run regressions in standard OLS form, and in an AR2 form to correct for second degree autocorrelation, for each of the three interpolation techniques being tested.
One potential criticism of the proposed method is that the estimates obtained via the control regressions are being held as truth, when in fact they are sample sensitive observations drawn from the governing parameter distribution\textsuperscript{xii,\textsuperscript{xiii}. In order to address this concern, in addition to the FRED data, four simulated data sets will be created so that regression results based on interpolated data can be held to known population parameters. These datasets will be constructed in a way so as to reflect the general conditions of the FRED samples in terms of the estimates’ magnitudes, though are considered as population datasets\textsuperscript{xiv}. Two of these datasets will be designed to be well-behaved, in that they will not suffer from autocorrelation. The first of these two data sets will contain 100 observations and the second 200 observation; both will be used to run standard OLS regressions. The second two datasets will be generated containing second degree autocorrelation, in order to examine the sensitivity of results based on interpolated data to autocorrelation correcting transformations. As is the case with the authentic data, the simulated datasets will contain 100 and 200 observations, each of which will be used to simulate situations in which 5% and 10% of the observations are missing.

To summarize, for both the artificial and authentic data, samples of 100 and 200 will be used to simulate 5% and 10% of the dependent variable observations as missing.

\textsuperscript{xii} It is standard practice in the profession to place some validity on the results obtained through regression analysis, despite this concern. Additionally, under the assumption of proper model specification the expected value of the estimated is equal to the true value, so results from the control regressions based on the data from the FRED database would still be valid.

\textsuperscript{xiii} This is made very clear by examining the differences in the parameter estimate for income between the 100 and 200 observation results.

\textsuperscript{xiv} It is important to note that all four of these data sets will be generated to reflect the characteristics of the data obtained from the FRED database when regressed in a Keynesian consumption function, though could technically reflect any arbitrary two variable time series model.
Furthermore, each of these combinations will be examined in a standard OLS and an AR2 model. The end result is that each type of data (authentic and artificial) will be used for 8 combinations regressions, based on model specification, size of the dataset, and the relative percentage of the dataset simulated as missing. All three interpolation techniques will be applied to each of these combinations.

To conclude this section the reader should be reminded that the goal of this research is to determine which interpolation method best restores the estimator to the value obtained from the full sample, and *not* to improve upon the OLS estimator itself.
CHAPTER IV
RESULTS AND DISCUSSION

The following sections contain a presentation of the results for the authentic and artificial data individually. It begins with the control regression results for the authentic data and general summary statistics for the frequency distributions of parameter estimates obtained from the 1,000 regressions. Next, results are presented graphically and summarized in tables demonstrating the quantity of estimates based on interpolated value that fall within (±) one-half the value of the standard error from the control regression estimates. This same organization is then applied for the simulated datasets. For consideration of the length of the body of this document only selected graphical results are presented in this chapter, while those remaining are left for the appendices.

Results for Authentic Data

Table 4.1 below contains a summary of the results for all control regressions performed on the authentic data obtained from the FRED database. Most notable is the dramatic difference between the intercept terms when the dataset increases from 100 to 200 observations. This, combined with the larger slope coefficients for the 100 observation datasets indicate that the propensity to spend income has increased over time, a result consistent with general economic theory. Additionally, the DW estimates indicate that the AR2 specification corrects for the high degree of autocorrelation in the
data. While this transformation has no significant effect on the slope coefficients, it does depress the magnitude of the intercept

Table 4.1: Control Regression Results, Authentic Data

<table>
<thead>
<tr>
<th></th>
<th>100 Observations</th>
<th>100 Observations</th>
<th>200 Observations</th>
<th>200 Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS Results</td>
<td>AR2 Results</td>
<td>OLS Results</td>
<td>AR2 Results</td>
</tr>
<tr>
<td>Intercept</td>
<td>-339.638</td>
<td>-328.9917</td>
<td>-81.0002</td>
<td>-58.5332</td>
</tr>
<tr>
<td></td>
<td>(14.6245)</td>
<td>(30.2339)</td>
<td>(8.9469)</td>
<td>(39.1593)</td>
</tr>
<tr>
<td>Income</td>
<td>.9843</td>
<td>.9831</td>
<td>.9415</td>
<td>.9409</td>
</tr>
<tr>
<td></td>
<td>(.0026)</td>
<td>(.0052)</td>
<td>(.0022)</td>
<td>(.0086)</td>
</tr>
<tr>
<td>N</td>
<td>100</td>
<td>100</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>DW</td>
<td>.9311</td>
<td>1.9158</td>
<td>.176</td>
<td>2.2125</td>
</tr>
<tr>
<td>R²</td>
<td>.9993</td>
<td>.9995</td>
<td>.9989</td>
<td>.9998</td>
</tr>
</tbody>
</table>

Standard errors in parenthesis
Significance levels: .1* .05** .01***

Tables 4.2 and 4.3 below contain the mean values and standard deviations of the 1,000 parameter estimates obtained through the repeated sampling. These results indicate that both the Reduced Sample and First Order methods produce very accurate slope estimates on average. Furthermore, the Modified Zero Order method produces a consistent downward bias in the estimates and a dramatically higher standard deviation for the distributions.
Table 4.2: Summary Statistics for Frequency Distributions Based on Authentic Data, OLS Specification

<table>
<thead>
<tr>
<th>Observation Dataset</th>
<th>Reduced Sample Method</th>
<th>Modified Zero Order Method</th>
<th>First Order Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>STD</td>
<td>Mean</td>
</tr>
<tr>
<td>100</td>
<td>.9842</td>
<td>(.0007)</td>
<td>.9489</td>
</tr>
<tr>
<td>5% Missing</td>
<td>Mean</td>
<td>(.0007)</td>
<td>.9489</td>
</tr>
<tr>
<td>10% Missing</td>
<td>Mean</td>
<td>(.0007)</td>
<td>.9489</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>(.0007)</td>
<td>.9489</td>
</tr>
<tr>
<td>200</td>
<td>.9832</td>
<td>(.0012)</td>
<td>.9415</td>
</tr>
<tr>
<td>5% Missing</td>
<td>Mean</td>
<td>(.0006)</td>
<td>.9415</td>
</tr>
<tr>
<td>10% Missing</td>
<td>Mean</td>
<td>(.0006)</td>
<td>.9415</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>(.0006)</td>
<td>.9415</td>
</tr>
</tbody>
</table>

Table 4.3: Summary Statistics for Frequency Distributions Based on Authentic Data, AR2 Specification

<table>
<thead>
<tr>
<th>Observation Dataset</th>
<th>Reduced Sample Method</th>
<th>Modified Zero Order Method</th>
<th>First Order Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>STD</td>
<td>Mean</td>
</tr>
<tr>
<td>100</td>
<td>.9831</td>
<td>(.0007)</td>
<td>.9489</td>
</tr>
<tr>
<td>5% Missing</td>
<td>Mean</td>
<td>(.0007)</td>
<td>.9489</td>
</tr>
<tr>
<td>10% Missing</td>
<td>Mean</td>
<td>(.0007)</td>
<td>.9489</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>(.0007)</td>
<td>.9489</td>
</tr>
<tr>
<td>200</td>
<td>.9415</td>
<td>(.0011)</td>
<td>.898</td>
</tr>
<tr>
<td>5% Missing</td>
<td>Mean</td>
<td>(.0011)</td>
<td>.898</td>
</tr>
<tr>
<td>10% Missing</td>
<td>Mean</td>
<td>(.0011)</td>
<td>.898</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>(.0011)</td>
<td>.898</td>
</tr>
</tbody>
</table>

No results are presented for the frequencies of the intercept values, as they are mainly important in forecasting. However, it should be noted that there is a similar result: the Reduced Sample and First Order methods produce distributions that seem to center around the control regression values while the Modified Zero Order Method produces a severe downward bias.

Simply by examining Tables 4.1-4.3 it appears clear that the Modified Zero Order method is a bad choice for continuing research when dealing with missing dependent variable observations. Not only does it produce a biased slope results (and an incredibly
biased intercept value), but it also varies more greatly in its predictions as indicated
through the distributional standard deviations. Since the bias produced by the Modified
Zero Order method is large compared to the standard error of the control regression
parameter estimates, graphical representations of these results only demonstrate a large
piling up of estimates outside of the control regression parameter distribution. As such,
graphical results for the Modified Zero Order method are omitted from the body of this
document and are left for the appendices.

Eliminating the Modified Zero Order method leaves the Reduced Sample and
First Order methods, which up until this point seem to be nearly identical in their
predictions. Unfortunately, while this simple table method of presenting the results is
useful for finding obvious biases, it makes it difficult to compare the Reduced Sample
and First Order methods, as they seem to produce nearly identical results on average.
The following graphical presentation of results in this section, combined with the
summary tables will indicate more clearly which method between these two should be
preferred.

Figures 4.1 and 4.2 shown below present frequency distributions for the
Reduced Sample and First Order methods respectively, created from regressions based on
100 observations with 5% values simulated as missing. Not unexpectedly, these
distributions are quite similar. Both distributions center on the control regression
parameter, and both appear to be slightly negatively skewed, but the First Order method
produces over-predictions more frequently than the Reduced Sample method.
Figure 4.1: Frequency Distribution for Reduced Sample Method: Authentic Data, 100 Observations, 5% Simulated as Missing, OLS Specification

Figure 4.2: Frequency Distribution for First Order Method: Authentic Data, 100 Observations, 5% Simulated as Missing, OLS Specification
Figures 4.3 and 4.4 shown below contain the frequency distributions for the AR2 specification based on a dataset with 200 observations and 10% simulated as missing. These results show a dramatic difference between the results for the Reduced Sample and First Order methods. The Reduced Sample method produces a well behaved distribution which centers only slightly above the control regression value, and has only a small tail of values greater than the control.

The First Order method yields a significant bias in comparison, as a bulk of the estimates produced by this method fall well above the control regression value. Furthermore, this distribution appears bi-modal if not tri-modal, as there is clearly a second clustering of values right around the control regression value, as well as a set of outliers that occur well below the control.

Figure 4.3: Frequency Distribution for Reduced Sample Method: Authentic Data, 200 Observations, 10% Simulated as Missing, AR2 Specification
To conclude this discussion of result based on the authentic datasets, Tables 4.4 and 4.5 below contain a summary the efficiency of all three methods. The Modified Zero Order method is included to demonstrate the dramatic effect of the bias that is produces in the estimates.

Table 4.4, which contains results for the OLS specification, demonstrates several interesting patterns. Most notably is the dismal performance of the Modified Zero Order method, which at its best barely produces 10% (101 out of 1,000) of its values within the desired interval. Quite oppositely, the Reduced Sample and First order methods produce very efficient results. When dealing with a small percentage of the data as missing, these two methods both produce over 90% of the results within the desired interval around the control estimate. However, these results show reduced reliability when dealing with a larger percentage of the data simulated as missing.
Table 4.4: Summary of Efficiency: Percent of Estimates within ( + / - ) One-Half the Standard Error from Control Estimate, Authentic Data, OLS Specification

<table>
<thead>
<tr>
<th></th>
<th>Reduced Sample Method</th>
<th>Modified Zero Order Method</th>
<th>First Order Method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>100 Observation Dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Missing</td>
<td>93.2%</td>
<td>10.1%</td>
<td>92.7%</td>
</tr>
<tr>
<td>10% Missing</td>
<td>82.5%</td>
<td>8.1%</td>
<td>80.1%</td>
</tr>
<tr>
<td><strong>200 Observation Dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Missing</td>
<td>93.3%</td>
<td>0%</td>
<td>91.7%</td>
</tr>
<tr>
<td>10% Missing</td>
<td>73.2%</td>
<td>0%</td>
<td>76.3%</td>
</tr>
</tbody>
</table>

Table 4.5 contains results for the AR2 specification. Similar to the OLS results, the Modified Zero Order method performs poorly, producing roughly 10% of its values within the desired interval at its best.

The results for the Reduced Sample and First Order methods differ in their AR2 specification from the OLS specification in two ways. First, it is clear that they no longer share similar degrees of efficiency as the Reduced Sample Method consistently outperforms the First Order method by much larger margins than was the case in the OLS specification. Secondly, the reduction in efficiency coming from a larger percentage of the data simulated as missing disappears for the Reduced Sample method. In its poorest performance this method yields 950 estimates, or 95%, within the desired band, while at its best it produces 997, or 97%. 
Table 4.5: Summary of Efficiency: Percent of Estimates within (+ / -) One-Half the Standard Error from Control Estimate, Authentic Data, AR2 Specification

<table>
<thead>
<tr>
<th></th>
<th>Reduced Sample Method</th>
<th>Modified Zero Order Method</th>
<th>First Order Method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>100 Observation Dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Missing</td>
<td>99.7%</td>
<td>11.1%</td>
<td>97.5%</td>
</tr>
<tr>
<td>10% Missing</td>
<td>97.5%</td>
<td>8.9%</td>
<td>94.5%</td>
</tr>
<tr>
<td><strong>200 Observation Dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Missing</td>
<td>97.7%</td>
<td>.6%</td>
<td>90.1%</td>
</tr>
<tr>
<td>10% Missing</td>
<td>95%</td>
<td>0%</td>
<td>69.2%</td>
</tr>
</tbody>
</table>

These results indicate that the results and efficiency of these estimators are sensitive to the specification of the model, and correcting for autocorrelation. The results for the AR2 model specification should carry more weight as this method represents a more proper specification of the model, tentatively suggesting that the Reduced Sample method is superior.

Results for Simulated Data

Table 4.6 below contains results for all control regressions based on the simulated data. As was previously discussed, these datasets were designed specifically to reproduce values for the intercept and slope estimates similar to those of the authentic data. Additionally, the reader should be reminded that the datasets tested with OLS regression were deliberately created to have low autocorrelation so the Durbin-Watson statistics near 2 are expected for the OLS Specifications, and the DW estimates for the AR2 specifications are the values after correcting for 2nd degree autocorrelation.

The main statistical difference between results based on this artificial data and results discussed above are a decrease in the standard errors for the intercept terms and an
increase in the standard errors for the slope estimates. However, these differences, such as lower standard errors on the intercept estimates, are simply artifacts of the data generation process and have no bearing on the validity of the datasets. The only real importance that these values play in fact, is in the efficiency method developed for this paper, which looks at an interval (+/-) one-half the value of these standard error terms from their respective parameter estimates.

Table 4.6: Control Regression Results, Simulated Data

<table>
<thead>
<tr>
<th></th>
<th>100 Observations</th>
<th>100 Observations</th>
<th>200 Observations</th>
<th>200 Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS Results</td>
<td>AR2 Results</td>
<td>OLS Results</td>
<td>AR2 Results</td>
</tr>
<tr>
<td>Intercept</td>
<td>-335.5453 ***</td>
<td>-327.4951 ***</td>
<td>-74.1822 ***</td>
<td>-61.0553 ***</td>
</tr>
<tr>
<td></td>
<td>(3.7831)</td>
<td>(5.9953)</td>
<td>(1.3343)</td>
<td>(5.9874)</td>
</tr>
<tr>
<td>Income</td>
<td>.9783 ***</td>
<td>.9821 ***</td>
<td>.9493 ***</td>
<td>.9476 ***</td>
</tr>
<tr>
<td></td>
<td>(.0049)</td>
<td>(.0091)</td>
<td>(.0011)</td>
<td>(.0041)</td>
</tr>
<tr>
<td>N</td>
<td>.100</td>
<td>100</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>DW</td>
<td>1.8931</td>
<td>2.0006</td>
<td>1.9557</td>
<td>2.0383</td>
</tr>
<tr>
<td>R2</td>
<td>.9974</td>
<td>.9966</td>
<td>.9997</td>
<td>.9997</td>
</tr>
</tbody>
</table>

Standard errors in parenthesis
Significance levels: .1* .05** .01***

Tables 4.7 and 4.8 below contain the mean and standard deviations values for the frequency distributions based on the simulated data. Similar to the results discussed above, these tables indicate that the Reduced Sample and First Order methods produce average values that coincide very strongly with the control results while the Modified Zero Order method consistently under-predicts. Additionally, these results indicate much higher standard deviations for all distributions based on the Modified Zero Order method.
Table 4.7: Summary Statistics for Frequency Distributions Based on Simulated Data, OLS Specification

<table>
<thead>
<tr>
<th></th>
<th>Reduced Sample Method</th>
<th>Modified Zero Order Method</th>
<th>First Order Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>STD</td>
<td>Mean</td>
</tr>
<tr>
<td>100 Observation Dataset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Missing</td>
<td>.9737</td>
<td>(.0012)</td>
<td>.9349</td>
</tr>
<tr>
<td>10% Missing</td>
<td>.9736</td>
<td>(.0017)</td>
<td>.888</td>
</tr>
<tr>
<td>200 Observation Dataset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Missing</td>
<td>.9493</td>
<td>(.0003)</td>
<td>.9062</td>
</tr>
<tr>
<td>10% Missing</td>
<td>.9493</td>
<td>(.0004)</td>
<td>.8591</td>
</tr>
</tbody>
</table>

Exercising Tables 4.6-4.8 makes it once again clear that the Modified Zero Order method is a poor choice for dealing with missing dependent variable observations. As was the case with the authentic data, this method produces on average a large downward bias on the slope estimates, coupled with a larger standard deviation. Additionally, it also produces an enormous downward bias on the intercept term. In light of this poor performance graphical results for this method are included only in the appendix. For consistency, the same combinations of model, size of dataset, and percentage of data simulated as missing will be presented for the simulated data as was for the authentic.
Figures 4.5 and 4.6 below depict the frequency distributions for the simulated data, based on a 100 observation dataset with 5% of the values simulated as missing. In both cases the results show a strong centering tendency at the control regression value. Interestingly, both distributions now show a slight positive skew, which is more pronounced in the Reduced Sample than the First Order method. However, these values represent only a small portion of the total observations and hence it is still reasonable to expect high levels of efficiency for both of these methods when applied to this data.

![Frequency Distribution of Parameter Estimates](image)

Figure 4.5: Frequency Distribution for Reduced Sample Method: Simulated Data, 100 Observations, 5% Simulated as Missing, OLS Specification
Figures 4.7 and 4.8 present the frequency distributions for the simulated data regressed in the AR2 model specification for a 200 observation dataset and 10% of the observation simulated as missing. These results show a dramatic difference in the behavior of the two distributions. The Reduced Sample method has a strong centering tendency at or slightly above the control regression value, while the First Order method produces a weak centering tendency that is biased upward from the control value. These results suggest that the Reduced Sample method is both more efficient and less biased than the First Order method, as demonstrated by Figures 4.7 and 4.8.
Figure 4.7: Frequency Distribution for Reduced Sample Method: Simulated Data, 200 Observations, 10% Simulated as Missing, AR2 Specification

Figure 4.8: Frequency Distribution for First Order Method: Simulated Data, 200 Observations, 10% Simulated as Missing, AR2 Specification
Table 4.9 contains a summary of the results for regressions based on simulated data with a standard OLS specification. These results are highly consistent with the results seen above for the authentic data. First and most notable is the extremely poor performance of the Modified Zero Order method which produces only .7% of its values within the desired interval around the control regression parameter estimate. Furthermore, these results indicate that for both 100 and 200 observation datasets, the results are sensitive to the relative percent of the data that are simulated as missing, again implying diminished performance when dealing with more missing dependent variable observations.

Table 4.9: Summary of Efficiency: Percent of Estimates within (+/-) One-Half the Standard Error from Control Estimate, Simulated Data, OLS Specification

<table>
<thead>
<tr>
<th>Reduced Sample Method</th>
<th>Modified Zero Order Method</th>
<th>First Order Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 Observation Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Missing 96.4% .7% 95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10% Missing 85.9% 0% 84.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 Observation Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Missing 95.4% 0% 96.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10% Missing 86% 0% 85.5%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Turning to the AR2 specification, Table 4.10 shown below paints a rather pessimistic picture, especially in comparison to all of the previous results. As has been the case throughout, the Modified Zero Order method performs dismally. However, these results also indicate a surprising reduction in the efficiency of both the Reduced Sample and First Order methods. Furthermore there are large discrepancies between the two methods’ efficiency which was previously observed.
Table 4.10: Summary of Efficiency: Percent of Estimates within (+/-) one-half the Standard Error from Control Estimate, Simulated Data, AR2 Specification

<table>
<thead>
<tr>
<th>Reduced Sample Method</th>
<th>Modified Zero Order Method</th>
<th>First Order Method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>100 Observation Dataset</strong></td>
<td>5% Missing 83.8% 7.4% 57.1%</td>
<td>10% Missing 54.7% 0% 30.6%</td>
</tr>
<tr>
<td><strong>200 Observation Dataset</strong></td>
<td>5% Missing 87.4% .1% 56.9%</td>
<td>10% Missing 75.2% 0% 53.7%</td>
</tr>
</tbody>
</table>

Unlike the authentic data, both specifications of regression based on these simulated datasets must be considered equally; the simulated data tested with OLS were intentionally created to not have autocorrelation, thereby requiring the OLS technique, while the simulated data tested with AR2 were intentionally designed to have second degree autocorrelation, making the AR2 specification necessary. Setting aside the fact that the values were tailored to be similar to the Keynesian consumption results, these simulated datasets are technically representative of any two variable time series model. Since it possible to have an authentic time series dataset that does not suffer from autocorrelation, the results obtained in this investigation are therefore more general.

Regardless of the type of data being used, the model specification, the size of the dataset, and even in all but a few cases the relative percentage of the data simulated as missing, these results clearly indicate that the Reduced Sample method outperforms the other methods. Based on the graphical representation and the tables summarizing efficiency, it is easy to see that the Reduced Sample method is the least biased, and the most efficient / consistent technique.
In summation, the Reduced Sample method outperforms the other two methods in (nearly) all cases, and the only exceptions to this are negligible differences between this method and the First Order method. The First Order method appears to have little to no inherent bias and a high degree of efficiency in most all cases, though is ultimately less effective than the Reduced Sample. The Modified First Order method performs poorly in all circumstances and should therefore not be considered as a viable method to compensate for missing dependent variable observations.
CHAPTER V

CONCLUSIONS

This paper has relied on a computer based Monte Carlo experiment to simulate situations in which missing dependent variable observations occur at random in order to understand how researchers can best compensate for this problem. Results from these regressions have been compared to a full information control model to demonstrate the efficiency and biases that arise when performing regression analysis on interpolated values. This comparison leads to the conclusion that the Reduced Sample regression method should be the preferred choice for dealing with missing dependent variable observations.

The result obtained in this investigation demonstrate that estimates obtained through regression analysis are moderately sensitive to model specification and the relative percentage of the data that is missing, and to a lesser extent the size of the dataset. Regardless, these results strongly indicate that the Reduced Sample method outperforms both the Modified Zero Order and First Order methods in terms of bias and efficiency. This conclusion stands in contrast to previous concerns that the Reduced Sample may yield biased results, as its’ reorganization of the dataset creates the likelihood of introducing unnatural jumps in the dataset. Based on the results obtained, there is no empirical justification for this concern.
While this paper presents a robust empirical testing of these methods in certain contexts, it has several shortcomings which should be addressed. First, no truly small datasets were tested. While the results obtained here show occasional differences between datasets containing 100 and 200 observations, these results may be dramatically different if dealing with a smaller dataset. However, decreasing the size of the dataset increases the possibility of inducing a small sample bias, thereby calling into question the validity of the research regardless of missing observations.

Secondly, the relative percentages of the data sets that were simulated as missing were relatively small. A stress test of sorts, in which the number of missing values is increased beyond 10%, may yield different results. The reason for this is that when only a small number of observations are missing (assuming that they are random) the probability that they fall near each other is small. However, if the number increases, the probability that multiple missing observations occur adjacent to each other increases, thereby increasing the likelihood of a jump in the dataset.

While this investigation has taken an essential step towards developing a framework in which researchers can compensate for missing dependent variable observations, the subject is far from concluded. Speaking only in a dependent variable context, there must be additional research focusing on multivariate models. In multivariable models when proper model specification is less clear, the Modified Zero Order method may yet prove helpful.

Methodologies similar to the one developed in this paper could be applied to test how to best compensate for missing independent variable observations as well. While the focus of this paper was practical in nature, more complex maximum likelihood methods
of interpolation could also be tested, assuming they could be adopted for a similar Monte Carlo investigation.

Researchers in all fields should always seek to base empirical work on complete datasets. However, if all feasible data collection methods have been exhausted, this paper presents strong evidence that the Reduced Sample method of compensating for missing dependent variable observations will yield both unbiased and highly efficient results.


APPENDICES
APPENDIX A.

ADDITIONAL OUTPUT FOR AUTHENTIC DATA: OLS SPECIFICATION

Figure A.1: Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 5% Simulated as Missing, OLS Specification

- Frequency Distribution of Parameter Estimates
- Control Regression Parameter Distribution:
  - Mean = .9843
  - Standard Error = .0026
Figure A.2: Frequency Distribution for Reduced Sample: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification

Figure A.3: Frequency Distribution for Modified Zero Order: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification
Figure A.4: Frequency Distribution for First Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, OLS Specification

Control Regression
Parameter Distribution:
Mean = .9843
Standard Error = .0026

Figure A.5: Frequency Distribution for Reduced Sample Method: Authentic Data, 200 Observations, 5% Simulated as Missing, OLS Specification

Control Regression
Parameter Distribution:
Mean = .9414
Standard Error = .0022
Figure A.6: Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, OLS Specification

Control Regression
Parameter Distribution:
Mean = .9414
Standard Error = .0022

Figure A.7: Frequency Distribution for First Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, OLS Specification

Control Regression
Parameter Distribution:
Mean = .9414
Standard Error = .0022
Figure A.8: Frequency Distribution for Reduced Sample Method: Authentic Data, 200 Observations, 10% Simulated as Missing, OLS Specification

Figure A.9: Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 10% Simulated as Missing, OLS Specification
Figure A.10: Frequency Distribution for First Order Method: Authentic Data, 200 Observations, 10% Simulated as Missing, OLS Specification
APPENDIX B.

ADDITIONAL OUTPUT FOR AUTHENTIC DATA: AR2 SPECIFICATION

Figure B.1: Frequency Distribution for Reduced Sample Method: Authentic Data, 100 Observations, 5% Simulated as Missing, AR2 Specification
Figure B.2: Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 5% Simulated as Missing, AR2 Specification

Figure B.3: Frequency Distribution for First Order Method: Authentic Data, 100 Observations, 5% Simulated as Missing, AR2 Specification
Figure B.4: Frequency Distribution for Reduced Sample Method: Authentic Data, 100 Observations, 10% Simulated as Missing, AR2 Specification

Figure B.5: Frequency Distribution for Modified Zero Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, AR2 Specification
Figure B.6: Frequency Distribution for First Order Method: Authentic Data, 100 Observations, 10% Simulated as Missing, AR2 Specification

Figure B.7: Frequency Distribution for Reduced Sample Method: Authentic Data, 200 Observations, 5% Simulated as Missing, AR2 Specification
Figure B.8: Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, AR2 Specification

Figure B.9: Frequency Distribution for First Order Method: Authentic Data, 200 Observations, 5% Simulated as Missing, AR2 Specification
Figure B.10: Frequency Distribution for Modified Zero Order Method: Authentic Data, 200 Observations, 10% Simulated as Missing, AR2 Specification
APPENDIX C.

ADDITIONAL OUTPUT FOR SIMULATED DATA: OLS SPECIFICATON

Figure C.1: Frequency Distribution for Modified Zero Order Method: Simulated Data, 100 Observations, 5% Simulated as Missing, OLS Specification
Figure C.2: Frequency Distribution for Reduced Sample Method: Simulated Data, 100 Observations, 10% Simulated as Missing, OLS Specification

Figure C.3: Frequency Distribution for Modified Zero Order Method: Simulated Data, 100 Observations, 10% Simulated as Missing, OLS Specification
Figure C.4: Frequency Distribution for First Order Method: Simulated Data, 100 Observations, 10% Simulated as Missing, OLS Specification

Figure C.5: Frequency Distribution for Reduced Sample Method: Simulated Data, 200 Observations, 5% Simulated as Missing, OLS Specification
Figure C.6: Frequency Distribution for Modified Zero Order Method: Simulated Data, 200 Observations, 5% Simulated as Missing, OLS Specification

Figure C.7: Frequency Distribution for First Order Method: Simulated Data, 200 Observations, 5% Simulated as Missing, OLS Specification
Figure C.8: Frequency Distribution for Reduced Sample Method: Simulated Data, 200 Observations, 10% Simulated as Missing, OLS Specification

Figure C.9: Frequency Distribution for Modified Zero Order Method: Simulated Data, 200 Observations, 10% Simulated as Missing, OLS Specification
Figure C.10: Frequency Distribution for First Order Method: Simulated Data, 200 Observations, 10% Simulated as Missing, OLS Specification

- **Parameter Distribution:**
  - Mean = .9493
  - Standard Error = .0011
APPENDIX D.

ADDITIONAL OUTPUT FOR SIMULATED DATA: AR2 SPECIFICATON

Figure D.1: Frequency Distribution for Reduced Sample Method: Simulated Data, 100 Observations, 5% Simulated as Missing, AR2 Specification
Figure D.2: Frequency Distribution for Modified Zero Order Method: Simulated Data, 100 Observations, 5% Simulated as Missing, AR2 Specification

Figure D.3: Frequency Distribution for First Order Method: Simulated Data, 100 Observations, 5% Simulated as Missing, AR2 Specification
Figure D.4: Frequency Distribution for Reduced Sample Method: Simulated Data, 100 Observations, 10% Simulated as Missing, AR2 Specification

Figure D.5: Frequency Distribution for Modified Zero Order Method: Simulated Data, 100 Observations, 10% Simulated as Missing, AR2 Specification
Figure D.6: Frequency Distribution for First Order Method: Simulated Data, 100 Observations, 10% Simulated as Missing, AR2 Specification

Figure D.7: Frequency Distribution for Reduced Sample Method: Simulated Data, 200 Observations, 5% Simulated as Missing, AR2 Specification
Figure D.8: Frequency Distribution for Modified Zero Order Method: Simulated Data, 200 Observations, 5% Simulated as Missing, AR2 Specification

Figure D.9: Frequency Distribution for First Order Method: Simulated Data, 200 Observations, 5% Simulated as Missing, AR2 Specification
Figure D.10: Frequency Distribution for Modified Zero Order Method: Simulated Data, 200 Observations, 10% Simulated as Missing, AR2 Specification