Two Essays in Applied Econometrics

by

Anubhab Gupta

A Thesis Submitted to the Faculty of the

DEPARTMENT OF AGRICULTURAL AND RESOURCE ECONOMICS

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

The University of Arizona

2013

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permissions for extended quotation form or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interest of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Anubhab Gupta

Anubhab Gupta

APPROVAL BY THESIS DIRECTOR

This Thesis has been approved on the dates shown below:

Satheesh Aradhyula

Dr. Satheesh Aradhyula
Agricultural and Resource Economics

5/10/13

Date

ACKNOWLEDGEMENTS

I would like to thank Dr. Satheesh Aradhyula for his constant encouragement and support. His guidance and insights have been invaluable throughout this project. I would also like to thank Dr. George Frisvold, Dr. Bruce Beattie and Dr. Gary Thompson for having prolonged discussions, helpful suggestions and comments. I am grateful to Ms. Deborah S. Boettcher of Pima County Assessor's Office and Dr. Mary Kamerzell, Superintendent of Catalina Foothills School District for providing assistance with collection of data. A special thanks to Mr. Gan Jin for his help with the use of ArcGIS software and other suggestions. I would also thank Mr. Niladri Gomes for making corrections in the draft and designing its layout. A final thanks to all my colleagues and other staff members at AREC for being very supportive and helpful.

Two Essays in Applied Econometrics

Essay I: Public School Open Enrollment and Housing Capitalization

Essay II: Sub-sample Estimators of Big Data

Essay I: Public School Open Enrollment and Housing Capitalization

Contents

List of Figures

List of Tables

Abstract

While the standard hedonic analyses on housing prices attempt to evaluate the impact of school characteristics on house value capitalization, this paper is an attempt to evaluate the same based on increasing open enrollment numbers in better school districts at a district level. It estimates the standard models used in the literature and addresses the issue of bias due to unobserved heterogeneity. This paper also attempts to disentangle the effects of open enrollment by taking a difference model and removing the time invariant unobserved neighborhood characteristics. Also it focuses on houses in the school district boundaries and evaluates the effect at the *buffer zones* while controlling for other observed characteristics. I find that open enrollment significantly increases housing prices in outer buffer zone at a declining rate and have varying effects in the inner buffer zone and results depend on model specifications. Controlling for most of the possible observed and unobserved characteristics, open enrollment has non-linear effects on housing prices in the inner buffer zone.

Chapter 1

Introduction

Economic literature in real estate markets, especially on housing prices, show that single house prices are higher in better school districts, all else equal. A number of researchers have quantified the value of school quality, location in a better school district, and other neighborhood characteristics by applying the hedonic method developed by Rosen (1974). Most of the research has been focused in disentangling the effects of school characteristics from other neighborhood and environmental characteristics. Other studies have relied on cross-sectional identification of relationship between housing prices and variables that can be used as proxy for perceived school quality in different school districts.

With the advent of charter schools and the change of public policies on open enrollment in public school districts, it is interesting to evaluate the impact of schools on housing prices in and around a good school district. In 1993, the State of Arizona approved open enrollment in all school districts, contingent upon availability of classroom space (A.R.S. § 15-816.01). Arizona schools started enrolling students from other neighboring districts and the immediate question that arises is whether single house buyers continue to pay a premium for purchasing a house in the best school districts. Also, it is important to figure out whether the effect of open enrollment on single house prices has eroded over time. It would also be interesting to note what happens to the houses in school boundaries. With open enrollment, people might choose houses outside the boundaries of good school districts which are proximal to good schools and pay a lesser price. Hence open enrollment might have implications not only for housing prices in better school districts but also other neighboring districts.

In this study I have used housing sales data, house characteristics data and school districts data for Pima County, Arizona. This study is restricted to ten school districts in and around the Tucson Metropolitan Area using data from 2001 to 2012. Here, we use the standard approach of estimations using log-linear models which has been used mostly in the literature of hedonic analysis. To address the issue of controlling for unobserved heterogeneity this paper uses the difference approach which washes out all the time-invariant unobserved characteristics. Here I

also consider a buffer zone in and around Catalina Foothills School District (generally recognized as the premier school district in the Tucson metropolitan area) and evaluate the effect of open enrollment in that district on 1 mile inside and 2 miles outside the boundary other things constant (controlled for).

The paper is organized as follows: Chapter 2 gives a review of the existing literature on related work in hedonic models, Chapter 3 discusses to the construction of dataset with a brief description of the study area, Chapter 4 provides the variables description and the summary statistics including the introduction of the *buffer zones*, Chapter 5 gives the premise of the estimation procedure and methodology, Chapter 6 contains the results and their possible interpretations, while Chapter 7 concludes.

Chapter 2

Literature Review

2.1 Hedonic models of housing prices on school characteristics

The introduction of hedonic models of housing prices on school characteristics dates back to Oates' (1969) seminal paper. Oates showed using data from 53 northern New Jersey municipalities that there is a positive relationship between housing prices and school expenditures. There have been numerous studies following that where researchers tried to explore this relationship between school characteristics and housing prices and also how the prices are affected by location of schools. The most important feature of these studies involved in disentangling the effect of schools on housing prices from other implicit characteristics which determine the price of a house.

Bogart and Cromwell (2000) used data from Cleveland area to evaluate the effect of schools on housing prices. They compared the sale prices of houses on either side of school-district boundary where there were otherwise uniform taxes and other public services. In this way they attributed the differences in prices across boundaries to better schools. However, they didn't specifically determine which attributes of schools home buyers valued. Also since the areas are contiguous it might be possible that the variation in sale prices are not only due to difference in quality of schools but also some other unobserved neighborhood quality variation.

While Black's (1999) work is similar to Bogart and Cromwell, she controlled for neighborhood quality by replacing the vector of observed characteristics with a full set of boundary dummies that indicate houses which share a district boundary. All relevant neighborhood or house characteristics are usually not observed and hence it is not possible to model and control for all the unobserved neighborhood and house characteristics. Typically failing to control for these unobserved characteristics lead to omitted variable bias and biased parameter estimates. Black specifically mentions two biases: first those that vary at school district level, e.g. property tax rates and public good provision; second, omitted variable bias both within and across school district boundaries, such as differing neighborhood characteristics . She included that bound-

ary dummies account for any unobserved characteristics shared by houses on either side of the boundary. Hence, her resulting model looks like the following, where K_b is a vector of boundary dummies and also, she uses a standard log-linear form for estimating the hedonic price model:

$$\log(price_{iab}) = \alpha + X_{iab}^{'}\beta + K_b^{'}\phi + \gamma test_a + \epsilon_{iab} \tag{2.1}$$

Black's findings suggest that parents care and pay a premium for schools with better test scores, attendance rates and other unobserved school quality characteristics.

Other similar studies include Kane and Staiger (2006) who also use differences in housing prices along assignment zone boundaries to disentangle the effect of schools and other neighborhood characteristics. Using data from Mecklenburg County, North Carolina from 1994-2001, Kane and Staiger find significant differences in housing prices along school boundaries, implying that better schools positively impact housing values. Their study also suggests that the effect of schools on housing values operates through the characteristics of the population living in different neighborhoods and the quality of housing stock in the neighborhood. Kane and Staiger also use the standard log-linear form for estimation and their independent variables include elementary school characteristics measured by averaged test score for grades 3-5 over 1993-1999, distance to elementary schools, house characteristics, census tract characteristics, geographic fixed effects and fixed effects for year, month, middle school, high school, and municipality.

Weimer and Wolkoff (2001) evaluated the relationship of school performance and housing values using non-contiguous district and incorporation boundaries to identify the school effects. Using data from Monroe County, NY their study confirms the importance of school outputs on housing values controlling for other factors such as student body composition, high school characteristics and other public services. Apart from using a standard log-linear form, Weimer and Wolkoff use a multiplicative functional form that allows for separation of sale price into a quality adjusted quantity and a locationally determined price of housing. Their multiplicative functional form looks like the following:

$$\log(SalePrice) = \log(X_L\beta_L) + \log X_S\beta_S + \epsilon \tag{2.2}$$

where X_S includes the house characteristics and X_L include the location-related characteristics listed in neighborhood, town, elementary school, high school, and school tax panels.

Another related study by Downes and Zabel (2002) show similar results but using school level data and not at district level. Using data from American Hosuing Survey and Illinois School-Report Cards for Chicago from 1987-1991, they assign to each house school-level data for the closest school and show that school level variables are significantly better in estimating house values than district level data. In their study, Downes and Zabel include measures of input and measures of output in the house price regression and correct for school quality endogeneity by using instrumental variables. They have alleviated the problem of correlation of unobserved individual error component with observed regressors by differencing the data. Their results also

suggest that homeowners are valuing school outputs and not the inputs.

2.2 Related Hedonic Models of housing prices on other non-school environmental characteristics

While almost similar approaches have been used by most of the studies in the literature on housing prices the most concerning issue remains that how the effect of school can be disentangled from other neighborhood characteristics while evaluating its impact on housing prices. While Black, and Kane and Staiger used the differences approach on similar houses with district dummies to address this issue, Downes and Zabel use time series differences to handle the same. In my study, I use similar estimation procedures to remove the unobserved heterogeneity of neighborhood characteristics and other school characteristics. This allows to obtain unbiased estimates of parameters by using a log-linear specification on the naïve cross-sectional model by creating a difference model. Subsequently, I use the theoretical concept that a fixed effects model which results in within estimation given the panel nature of the dataset is no different from a difference model if only two time periods are considered. But first let us look at other related literature on hedonic models on housing prices, e.g. location within a certain geographical characteristics and other environmental settings.

In a related study on the effect of riparian habitat on housing prices in Tucson, Arizona by Bark et al. (2009), results show that homebuyers value quality riparian habitat and distinguish between biologically significant riparian vegetation characteristics. This hedonic study also uses a log-linear specification and controls for house and other neighborhood characteristics. They also disaggregate the environmental quality and show that buyers pay premiums for wetness and diversity and not for greenness per se. Bourne (2007) studied the effect of Santa Cruz riparian corridor, Rio Rico, Arizona on single family houses. She rejected a standard log-linear form by testing in favor of specification and ended using Liebig type distance equation for the econometric specification. Her study also suggests that within a certain distance from the riparian vegetation, homebuyers pay a premium for proximity to the riparian corridor. Hence, in related literature also there is evidence of homebuyers valuing certain house characteristics and paying a premium for the same. Also, it is empirically possible to disentangle the effect of some particular feature associated with a house from others by adopting certain econometric tools.

2.3 Open Enrollment in School Districts and housing prices

There are many studies which try to evaluate the effect of open enrollment in school districts, advent of charter schools, magnet schools and the expansion of private schools on several social issues, viz. parental decision making, difference in education deliverance from public schools, equity in the form of economic outcomes and other ethnic outcomes, and mobilization of homebuyers (Goldhaber, 1999), goals of integration and open enrollment (Smith, 1995), supply and demand

theory of educational choice (Funkhouser and Colopy, 1994), early effects of open-enrollment on significant changes in district open enrollments (Rubenstein, 1992). Other studies indicated that open enrollment resulted in transfer of students from one district to another based on higher student performance and higher socio-economic status from the districts they left (Fossey, 1994).

Seminal work by Reback (2005) was the first attempt to capture the effect of school choice program on housing prices. Using data from inter-district choices in Minnesota, Reback finds that residential properties appreciate significantly in those school districts where students are able to transfer while house values decline in those districts which accept transfer students. Reback uses percentage change in equalized, assessed value of all residential property in a school district between 1989-1990 and 1997-1998 as the dependent variable. This variable reflects annual changes in actual sale prices controlling for the fixed effects of differences in assessment practices. The right hand side variables include percentage change in number of students who transfer into a district, percentage change in number of students who transfer out of a district, a dummy for a district where no students have transferred and other house and resident characteristics. Reback uses a percent change in the dependent variable and hence it is a deviation from the standard log-linear form used in most hedonic analyses. He also controlled for changes in school quality by including the district level test score measure as an independent variable. For biases that might be present due to nonrandom sample attrition, he uses maximum-likelihood estimation of a Heckman (1976) selection model. His results suggest that both incoming and outgoing transfer rates have large statistically significant effects on the future growth rate of residential properties of a school district. The signs of the effects on those districts which accept incoming students have a significant decline in property values whilst the school districts which have outgoing students have an appreciation in assessed values of houses.

2.4 Contribution to literature

In this paper I explore a rich dataset available for Pima County, Arizona to assess the effect of open enrollment on housing prices using a log-linear specification, the standard model used in the literature on hedonic analysis. The first deviation from the most related work by Reback would be the use of actual prices and not assessed prices and hence I don't need to control for assessment practices which presumably would reduce any bias which was not controlled for. Second, I use the Case-Schiller price index to control for business cycles in the economy to eliminate that source of estimation bias. Third, a stark difference from Reback's work is that all school districts have open enrollment policy in place and hence no separate analysis is required for outgoing and incoming students in any particular district. Also the nature of the unbalanced panel dataset allows for houses which have been sold more than once in the study period and hence a difference model is evident. This resolves the issue of controlling for unobserved neighborhood characteristics as a first difference washes out the endogeneity of time invariant variables with other regressors in the model. A final sophistication is introduced by considering houses within a buffer zone

around the district boundaryof the best school district and a separate analysis is done for those houses which were sold in the buffer zone to evaluate the effect of open enrollment on houses which have almost similar characteristics separated by a school boundary. I also tested whether the parameter estimates are robust if the buffer zone is changed both within and outside the boundaries of the best school district.

Chapter 3

Description of Data

3.1 Study Area

The area studied in this paper is Pima County, Arizona which is the most populous county in southern Arizona. Pima County has 18 school districts which in total have 241 non-charter public schools. Figure 3.1 shows the Pima County school district map and the boundaries of all school districts. Out of these 18 school districts 14 are unified, 2 are transportation districts, 1 is an accommodation district and 1 is the Joint Technical Education District (JTED). All the school districts in Pima County have open enrollment policies. Arizona State (A.R.S. § 15-816.01) stipulates:

"School district governing boards shall establish policies and shall implement an open enrollment policy without charging tuition. A school district may give enrollment preference to and reserve capacity for pupils who are children of persons who are employed by or at a school in the school district. A copy of the district policies for open enrollment shall be posted on the district's website and shall be available to the public on request."

This means that with the open enrollment policies in place, parents have the choice of enrolling their school-aged children in a different school district than their district of residence.

For this study I considered only 10 school districts for the analysis as the effect of schools on the housing prices will be different for all the districts in Pima County and hence systematically 6 of the school districts were not considered (see below). The two other districts which were ignored in this analysis were the Pima Accommodation district and the Pima JTED as preferences for houses probably do not depend on these two districts.

The other six school districts which were not considered in the study area are: Ajo Unified School District (District 15), San Fernando Elementary School District (District 35), Empire Elementary School District (District 37), Continental Elementary School District (District 39), Indian-Oasis Baboquivari School District (District 40), and Redington Elementary School District (District 44). All of these are 'out-lying' districts (not part of Tucson Metropolitan Area).

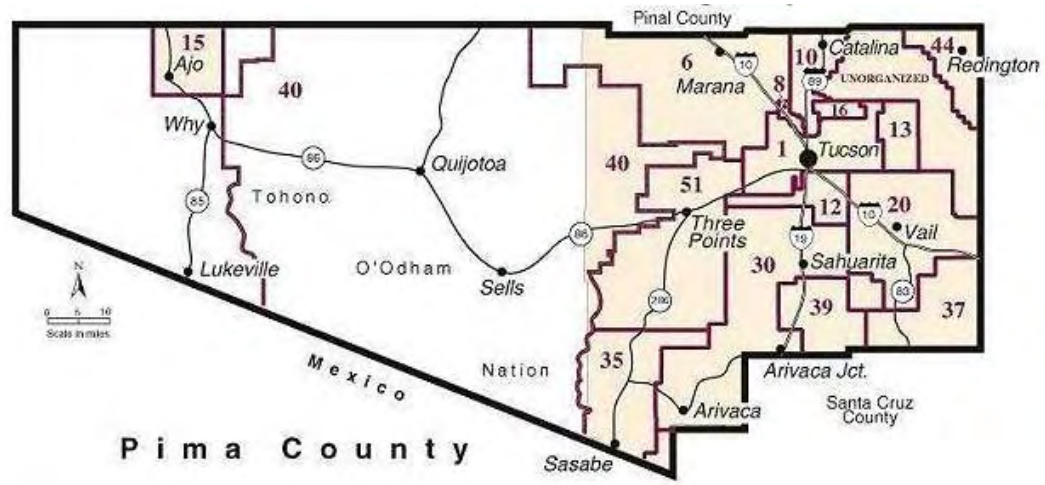

Figure 3.1: School District Map of Pima County, Arizona

Hence the focus of study narrows down to the districts in and around Tucson Metropolitan area. Figure 3.2 provides us with the map of school districts which has been considered in this paper.

The study area analyzed in this paper consists of the following 10 school districts: Tucson Unified School District (TUSD, Dist. 1), Marana Unified School District (Dist. 6), Flowing Wells Unified District (Dist. 8), Amphitheater Unified School District (AUSD, Dist. 10), Sunnyside Unified School District (Dist. 12), Tanque Verde Unified District (TVUD, Dist. 13), Catalina Foothills Unified District (CFSD, Dist. 16), Vail Unified School District (Dist. 20), Sahuarita Unified School District (Dist. 30), and Altar Valley Elementary District (Dist. 51).

Catalina Foothills School District is considered to be the best school district among all the others considered and the boundary of CFSD is marked in red in Fig. 2.2.

3.2 Description of Data Sources

3.2.1 House Characteristics by School Districts

The data on house characteristics come from Pima County Assessor's Office and are available on their website publicly. The data used here is in their library under the label Tax Year 2013. Three datasets taken from this folder were: Mas.zip, Notice.zip and EDNPI.zip. Mas.zip contains individual house's data on house characteristics, viz., living area (in sq. ft.), number of bedrooms, garage, year built etc. There are 279,332 observations in this dataset. The unique identifier is given by the variable "Parcel" and this variable is used for merging the other datasets with Mas.zip. Notice.zip has house level data on each house's address, their area code and has 441,635 observations. The area code variable in this dataset is used to assign each house to a school district which also falls within the same area code. Finally, EDNPI.zip contains information on latitudinal and longitudinal location and the total land area (in sq. ft.) of each

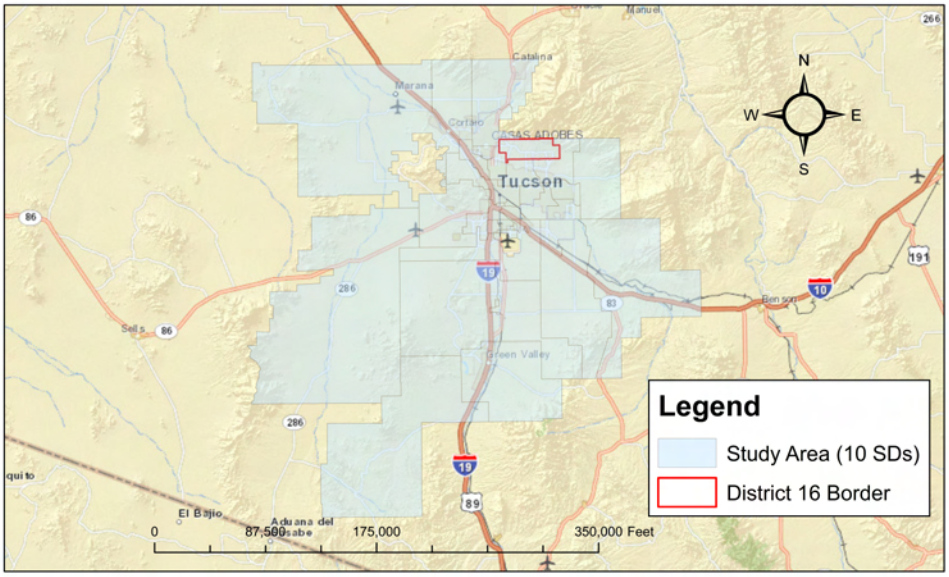

Figure 3.2: Study Area considered in Pima County

house. EDNPI.zip has 417,034 observations. These three datasets were merged using the unique identifier of each house, Parcel. Before merging the houses, the three datasets were cleaned and houses with missing Parcel IDs were deleted. After merging the three datasets we have 255,890 observations on the final merged dataset. But clearly all these houses were not sold even once within last 12 years.

3.2.2 Home Sales, Housing Price Index and Fixed Rate Mortgage

The sales data on all the houses in Pima County is collected for 12 years starting from 2001 to 2012 and this data is also available from Pima County Assessor's Office under the label Miscellaneous Data Files in their data library. The twelve zip files each representing sales data for each year were merged together and an unbalanced panel data of sales was created and cleaned. The cleaning included deleting houses without parcel Ids and duplicate sales data with same sale price within a month in a given year were also removed resulting in 188,584 observations.

The housing price index used for the analysis in this paper is the "Case-Schiller Price index (CSPI)" which is used as a standard index for normalizing sales prices. Data on monthly CSPI is available from Standard and Poor Dow Jones Indices website. Seasonally adjusted Home Price

Index levels as of January, 2013 was collected to account for seasonality in sales price and this was merged with the unbalanced panel of sales data using month of sales and sale year as a merging variable. CSPI used is the index for Phoenix, AZ metro area and is the best available to use as a proxy for Pima County.

Fixed rate mortgage (FRM) data comes from hsh.com and the monthly national average for 30 year FRM was collected and used as a proxy for the study area.

This unbalanced panel data on sales now with the CSPI and FRM was finally merged with the merged dataset available on house characteristics using Parcel Id as the unique identifier for each household.

3.2.3 Enrollment and Charter School Data

The open enrollment in CFSD is proprietary data which was available upon request from the CFSD's Superintendent. This data contains information not only on the district 16's open enrollment from 2001-2012 but also on total enrollment in District 16. The data on enrollment in charter schools comes from publicly available Arizona Department of Education's Research and Evaluation group's October 1st enrollment information for the entire state of Arizona. This contains charter enrollment data from 2006 onwards for the entire state of Arizona and this variable is used as a proxy for the study area.

3.3 Final Dataset

The final unbalanced panel dataset was obtained upon merging the data on house characteristics from Mas.zip, Notice.zip and EDNPI.zip; sales datasets for 12 years along with FRM and CSPI; and data on open enrollment in CFSD and charter schools enrollment. Also as indicated earlier, the final dataset contains "single houses" only in the 10 districts considered in the study area. This dataset was cleaned and resulted in 113,239 houses with 170,291 observations. Since some of the houses were sold more than once within this twelve years time period we have the unbalanced panel nature of this dataset. Table 3.1 gives the description of number of houses and how many times they were sold.

Number of times sold	Number of houses	Observations
1	69,873	69,873
2	32,117	64,234
3	9,137	27,411
4	1,814	7,256
5	273	1,365
6	23	138
7	2	14
TOTAL	113,239	170,291

Table 3.1: House Sales Information

Chapter 4

Variables and Summary Statistics

4.1 House Characteristics and School Districts

4.1.1 Description of Variables

The key variables that were included in the analysis on house characteristics and school districts are given in the Table 4.1.

The dummy variables on house characteristics were created additionally which included the dummies for pool, covered garage, patio, air conditioned cooling, evaporative cooling and the four dummies for assessed qualities given by the assessors' office. The unique identifier variable is a 9-digit unique number which has been used to merge and identify unique houses. The dummy variables on each school district have also been generated in order to capture the fixed effects of school districts in the analysis.

In this study I have also considered a buffer zone around the CFSD boundaries. This buffer zone was carefully constructed to identify those houses located in this zone. All houses in the dataset had information on their exact latitude and longitude, both measured in degrees and upto four decimal points. All the houses were plotted using ArcGIS software and were superimposed on the school district map from Pima County GIS website. The school district map from Pima County GIS is proprietary and needed special access for this study. After plotting the houses in the school district map, a two-mile stretch outside and a one-mile stretch inside the CFSD boundary was constructed and the houses only within this zone were identified. Fig 4.1 provides us with the generated map of the buffer zones. The boundary is shown by the black border and each house that falls within this buffer zone is given by a green dot. Since the north eastern boundary of CFSD is covered by the Catalina Mountains, we have not considered any houses in the north eastern buffer zone as they are not indicative of a boundary house. Also, there are no houses outside northern boundary of CFSD and hence all the houses north of 32.3076 degrees latitude and east of -110.9104 degrees longitude were systematically removed from the buffer zone. This leaves with houses in the outer buffer zone in Tucson Unified School District

Variable Name	Description
Parcel	Unique Identifier of each house (9-digit code)
Rooms	Total number of rooms in the house
SQFT	Livable Area (measured in square foot)
Landsqft	Total Area of the Plot (measured in square foot)
Bathfixtur	Number of total Bath Fixtures
d_pool	Dummy variable for pool
d_garage	Dummy variable for covered garage
d_patio	Dummy variable for patio
d_ac_cool	Dummy variable for air conditioned cooling
d_ev_cool	Dummy variable for evaporative cooling
d_min_quality	Dummy variable if assessed quality is minimum
d_fair_quality	Dummy variable if assessed quality is fair
d_good_qualtiy	Dummy variable if assessed quality is good
d_excellent_quality	Dummy variable if assessed quality is excellent
d_new	Dummy if house is sold within 5 years of construction
d_tucson	Dummy if house in Tucson Unified School District
d_marana	Dummy if house in Marana Unified School District
d_flowingwells	Dummy if house in Flowing Wells Unified District
d_amphitheather	Dummy if house in Amphitheater Unified School District
d_sunnyside	Dummy if house in Sunnyside Unified School District
d_tanque	Dummy if house in Tanque Verde Unified District
d_cat	Dummy if house in Catalina Foothills Unified District
d_vail	Dummy if house in Vail Unified School District
d_sahuarita	Dummy if house in Sahuarita Unified School District
d_altar	Dummy if house in Altar Valley Unified District

Table 4.1: Variable Description of House characteristics School Districts

from south and east, and Amphitheater School District in the west. Two separate dummies were constructed: *d_outer_buffer* and *d_inner_buffer* for a house located outside and inside the CFSD boundary respectively.

4.1.2 Summary Statistics

The summary of the key house characteristics variables are given in Table 4.2. Figure 4.2 shows how the number of observations from each school district.

4.2 Housing Sales and Open Enrollment

4.2.1 Description of Variables

The sales price for all the houses obtained from 2001-2012 were normalized by using the Case-Schiller Price Index (CSPI) to take into account the business cycles in the economy and to control for seasonality in prices. CSPI is the US national index for housing prices that tracks the value

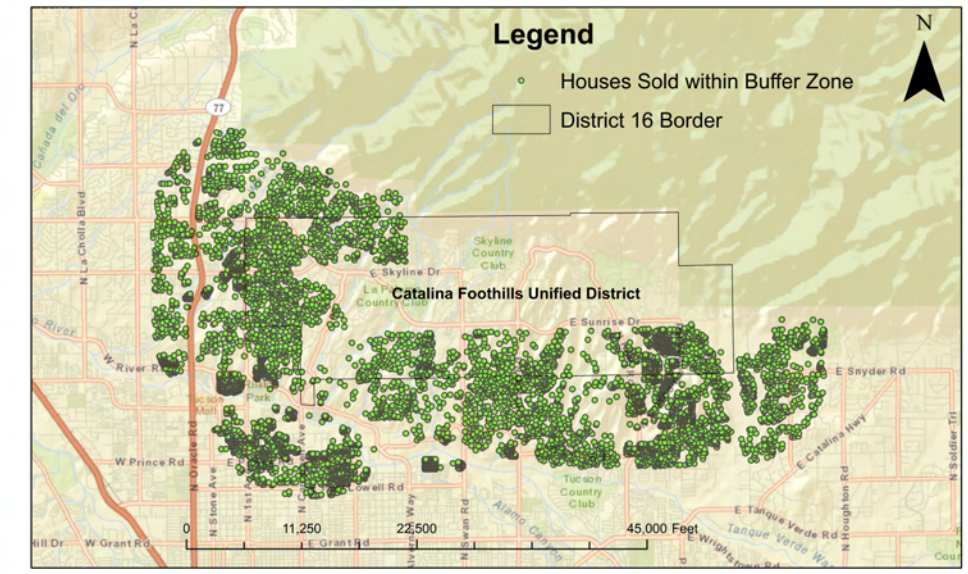

Figure 4.1: buffer zone in and around CFSD boundary

of single family housing within the U.S. and is calculated monthly using a three-month moving average algorithm. I used the following index for convenience of interpretation. This index is used for normalizing the sales price and is defined as the following:

$$\text{Index} = \frac{\text{CSPI}}{100} \tag{4.1}$$

Figure 4.3 provides the graph showing how the yearly median values of CSPI fluctuated in the last twelve years considered in my study.

The other variables of interest in this study are: the open enrollment in Catalina Foothills School District from 2001 onwards; and the charter school enrollment in the state of Arizona that has been used as a proxy for Pima County. Table 4.3 defines the sales prices and open enrollment variables.

4.2.2 Summary Statistics

Figure 4.4 plots the median sale prices, the normalized sale prices overall and the normalized sale prices in CFSD.

Variable	Mean	Min	Max	Standard Deviation
Rooms	6.9	1	50	1.5602
SQFT	1833.4	224	10901	662.0775
Landsqft	13279.2	140	2006163	26139.5
Bathfixtur	7.23	3	30	2.3812
d_pool	0.2058	0	1	0.4043
d_garage	0.7461	0	1	0.4353
d_patio	0.8215	0	1	0.3829
d_ac_cool	0.8186	0	1	0.3853
d_ev_cool	0.1813	0	1	0.3853
d_min_quality	0.0139	0	1	0.1170
d_fair_quality	0.5961	0	1	0.4907
d_good_quality	0.3868	0	1	0.4869
d_excellent_quality	0.0038	0	1	0.0613
d_new	0.4211	0	1	0.4937

Table 4.2: Summary Statistics on House Characteristics

Variable Name	Description
SalePrice	Sales Price of each house
Price	Sales Price normalized by the index given in equation 4.1
ln_Price	Natural logarithm of Price
Price_liv_area	Price per square feet of livable area
ln_Price_liv_area	Natural logarithm of Price_liv_area
Cat_Op_Enroll	Open enrollment numbers in CFSD in each year
Charter_Enroll	Yearly Charter School Enrollment in Arizona from 2006

Table 4.3: Description of Variables on Housing Sales and Open Enrollment

Table 4.4 shows the summary statistics of the key variables described in Table 4.3. Finally Figure 4.5 provides us with the total number of in-district students and open-enrolled students in CFSD.

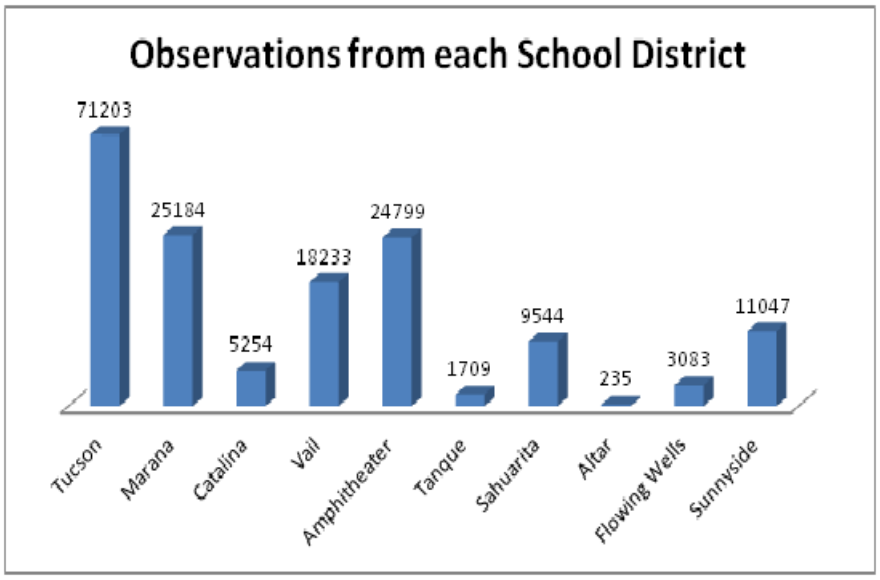

Figure 4.2: Total Number of Sales for all 10 school districts

Variable	Mean	Median	Min	Max	Standard Deviation
SalePrice	223,412.24	180,000	10,000	8,000,000	267,699.4
Price	158,428.81	124,634	4,532.48	7,643,799	235,780.1
ln_Price	11.787	11.733	8.419	15.85	0.533
Price_liv_area	83.75	75.11	2.428	6,306.76	128.884
ln_Price_liv_area	4.331	4.319	0.887	8.75	0.356
Cat_Op_Enroll	388.414	165	113	1408	400.496
Charter_Enroll	112,993.05	100,701	93,213	145,273	18,195.08

Table 4.4: Summary Statistics on variables in Table 4.3

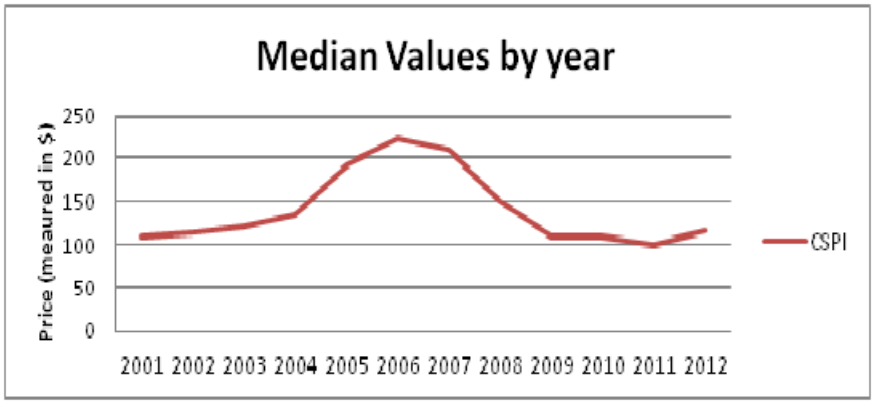

Figure 4.3: Case Schiller Price Index

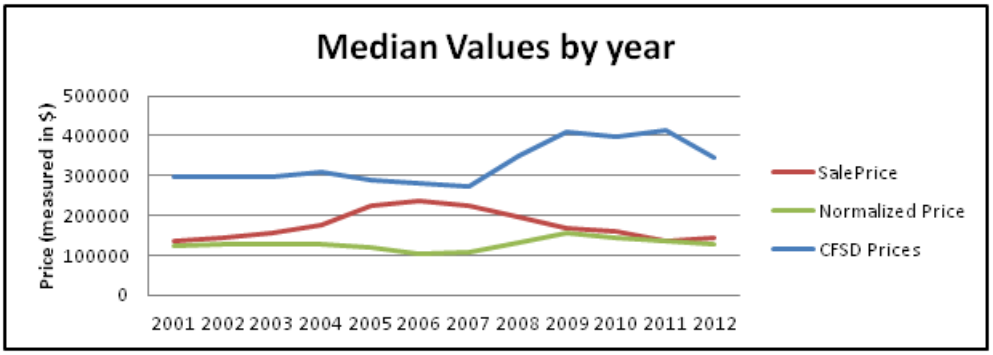

Figure 4.4: Median Sale Price, Normalized Sale Price (overall) & CFSD normalized Sale Prices by years

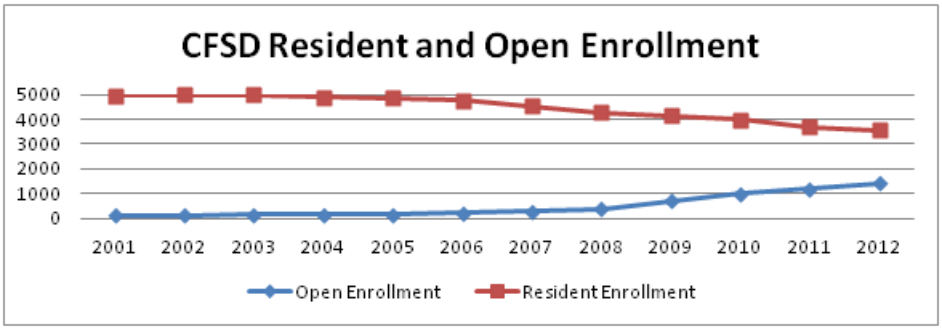

Figure 4.5: Resident and Open Enrollment in CFSD

Chapter 5

Empirical Models and Methodology

5.1 Conceptual Framework

The main approach of this paper is to identify the changes in single house values owing to a change in open enrollment numbers in these districts. For this purpose, I have considered how housing values have changed from 2001-2012 in the 10 districts of Pima County due to increase in open-enrollment numbers in Catalina Foothills School District (CFSD). While it is not possible to delineate the exact change of policy as all the school districts have adapted open enrollment regime slowly over this period it would not be possible to capture the pre and post effects of this policy change on housing prices, however, as we have open enrollment numbers in CFSD for the entire time period, we might attribute changes in housing prices owing to increase in open enrollment over the years.

The critical challenge for this kind of a hedonic analysis is to disentangle the effects of school districts open enrollment policies from neighborhood and other unobserved characteristics. Also since in this study we are restricted at the district level effects, I do not need to control for every school characteristics as we might safely assume that school effects matter within a district but the district is representative of all the school characteristics. Hence by using dummies to control for, we will be able to capture most of the school characteristics in our analysis at the district level.

A hedonic regression analysis model for estimating the value of residential houses on various attributes of housing properties has the general form:

$$SalePrice_{it} = f(X_i, D_i, N_i, FRM_t, D_i * OE_t, u_{i,t}) \tag{5.1}$$

where, $i = 1(1)N$, $t = 1(1)T$

$SalePrice_{it}$: Sales Price of House i in time t

X_i : Vector of time invariant House i's characteristics

D_i : Vector of School District dummy for each i

FRM_t : Fixed Rate Mortgage in time t

N_i : Vector of time invariant neighborhood dummy for each i

$D_i * OE_t$: District dummy interacted with Open Enrollment (OE) in time t

$u_{i,t}$: Measurement Error for each house in each time period

In this paper I have considered two approaches: one which has been mostly used in the literature for estimation of hedonic models on housing prices; second is a difference model which has been adapted to delineate the effects of school district open enrollment from other unobserved characteristics which are difficult to control for otherwise and hence might lead to biased estimates in the first approach. We can use the second approach because this dataset has many houses which were sold more than once in the time period of our study.

5.2 Standard Log-linear Model

The log-linear approach has been a standard procedure of estimation used in the housing price literature to estimate housing prices. This approach involves estimation of log of prices on all the possible right hand side variables which, for our purposes would include the house characteristic variables, school district dummies and other variables. In this approach, we will consider our unbalanced panel data as a cross-sectional data and analyze. This in effect means that each house which has been sold more than once in this time period will be considered as separate observations and the panel nature of the dataset be ignored in this analysis. Hence the basic framework of regression model for this paper would like the following:

$$\log(price_{it}) = \alpha + X_i'\beta_1 + \beta_2 FRM_t + D_i'\beta_3 + (D_i * OE_t)'\beta_4 + u_{it} \tag{5.2}$$

where, $i = 1(1)N$, $t = 1(1)T$ and $u_{it} \sim N(0, \sigma_t^2)$, and all variables are defined as above.

Also of additional interest to us would be the effect of open enrollment around the borders of CFSD, i.e., the effect on the buffer zone given in Figure 4.1. Hence we interact open enrollment

numbers on the dummies for outer buffer zone and inner buffer zone. From equation (5.2) it is quite clear that the vector of house characteristics, district and buffer dummies are time invariant while the 30 year fixed rate mortgage, dummies interacted with open enrollment numbers and the unobserved errors are time varying and some of them vary across cross-section.

5.3 Difference Model

The panel nature of the dataset has been used to exploit the possible removal of unobserved heterogeneity across households by the implementation of a difference model. The underlying assumption is that the house characteristics and other observed and unobserved characteristics which don't change over time get washed away when a difference is taken of the same house in a different time period with itself. Hence the contributing factors to the change of sale prices would thus be attributable to only the time-varying characteristics. Some of these variables might also be varying across cross-section but the essence of such an approach removes the unobserved characteristics which cannot otherwise be controlled for.

Hence the framework of a difference approach in a log-linear model would be the following:

$$\log\left(\frac{price_{it}}{price_{i,t-k_i}}\right) = \beta_2(FRM_t - FRM_{t-k_i}) + [(D_i * OE_t) - (D_i * OE_{t-k_i})]'\,\beta_4 + (u_{it} - u_{i,t-k_i}) \tag{5.3}$$

where $i = 1(1)N$, $t = 1(1)T$;
$k_i = 1(1)11$ is the difference between the years of sale for each house;
$\epsilon_{it} = (u_{it} - u_{i,t-k_i}) \sim N(0, \sigma_t^2 + \sigma_{t-k_i}^2 - 2\text{cov}(u_{it}, u_{i,t-k_i}))$.

One important correction that needs to be made because k_i varies for each house if it has been sold more than once. So in order to control for these differences, I include the difference in years of sale for each house as a right hand side variable in equation (5.3). Finally in my difference model, I also include dummies for each year of last sale to capture for any other correction due to time trends.

As given in Table 3.1, this dataset has 32,117 houses which were sold twice and 9,137 houses which were sold thrice in between 2001-2012. I considered only these 41,254 houses for our difference model. For houses which were sold thrice, I took the difference between the third sale and the first sale for constructing single differences. Hence all these houses in the difference model will have a single observation which would be a differenced value for the normalized price and all the time-varying right hand side variables. The time-invariant variables as suggested would drop out.

Chapter 6

Results and Implications

6.1 Regression Results

As discussed in Chapter 5, I have considered two basic frameworks for estimating housing prices and house capitalization. The first is the standard log-linear model which has been used in most of the studies so far. The second approach is an attempt to remove possible unobserved heterogeneity across houses in different school districts. This approach as described in the previous chapter involves taking difference of those houses which were sold more than once. Regression results of the eight estimated models are presented in the appendix. Four of these models are done using the log-linear framework and the remaining four uses a first difference of the natural logarithms of the dependent variables used in the first four models respectively. Hence, as defined, the last four are the difference models. A brief discussion of the parameter estimates are given in the next subsection. The only variant in the first two models is that in the first log-linear model I have used natural logarithm of normalized prices as the dependent variable and in the second one a natural logarithm of normalized prices per square footage of livable area as the dependent variable using all observations. The third and the fourth models have the same respective dependent variables but with observations from 2006 onwards as we can now use the data available on Charter Enrollments and see their effects when interacted with district and buffer dummies while controlling for them. All the estimates reported are heteroscedasticity consistent and standard errors are reported in parenthesis with the corresponding p-values.

6.1.1 Parameter Estimates of Log-linear Models

Let us first consider the standard log-linear results given in the appendix. These models include natural logarithms of normalized prices regressed on all the house characteristics, FRM, yearly dummies, district dummies and buffer dummies, and some interesting interactions of Open Enrollment (OE) in CFSD with the district and buffer dummies. I have used Altar Valley School District and the year 2001 as reference.

In Model I, all the parameter estimates have expected signs. Here, I do not discuss the estimates for all the variables we control for and thus focus on the variables of interest. Interesting and intuitive estimates are those on the interaction of OE in CFSD with the outer buffer dummy. It has a positive and significant coefficient of 0.0002063 suggesting that with increase in OE, prices have increased in the outer buffer zone and has decreased after a certain number of open enrollments as the quadratic coefficient is 6.19×10^{-8}. Also, this model suggests that OE doesn't have any significant effect on the housing prices in the inner buffer zone since both the linear and squared terms are statistically insignificant. But, a house which is in the inner zone is also in CFSD and hence just the parameter estimates are not representative as we have to account for the OE interaction with CFSD. Similar is true for outer buffer zone which is stretched out in AUSD and TUSD. It is also interesting to note that OE when interacted with dummy for TUSD gives a negative and significant coefficient of 6.06×10^{-5}. This is difficult to interpret and one possible intuition could be that with OE, home buyers are willing to pay higher to be on the northern boundaries of TUSD rather than in the other parts of the district. Possibly, the higher demands for houses on the boundaries of District No.1 are driving the prices down elsewhere in the district.

Model II has the same regressors with log of normalized prices per square footage of livable area as the dependent variable. This model gives similar results with almost all the interesting interaction variables including the one with OE and outer buffer zone. We also find similar results as in Model I with OE interacted with dummy for TUSD. But in this model we find that OE has a positive and significant effect on the inner buffer zone and that effect is negative after a certain number of open enrollments. This result is starkly different from Model I and hence needs further attention.

Model III and Model IV provide regression results when I restricted the dataset to observations from 2006 onwards. The dependent variables are same as in the first two models respectively. While all the parameter estimates have expected signs, most of these are insignificant. For Model III, I obtain a 5% significant and positive estimate on the outer buffer zone when interacted with OE in CFSD while a negative coefficient on the effect on inner buffer zone which is significant at 10%. The effects on the non-linear terms for both these variables are not significant showing expected linear effects on both the zones. Also we find that Charter Enrollments (CE) has a negative effect on the outer buffer zone while has positive effects on inner zone and the remaining district dummies. Model IV is similar to Model II but with observations from 2006 and we find almost all insignificant impacts on key variables at 5% level of significance. Still effect of OE in TUSD is negative and for the possible interpretation given above.

6.1.2 Parameter Estimates of Difference Models

By invoking the difference model we can wash away all the time invariant house characteristics and the fixed effects of district dummies. Also it removes all the unobserved neighborhood characteristics which are fixed over time and thus gives a model where the difference in logarithms

of prices is just a function of the difference of time varying variables. I construct the difference models based on the first four standard log-linear models. Hence Models V, VI, VII and VIII are the differenced versions of Models I through IV respectively.

The parameter estimate on difference is sales years has a significant negative sign suggesting that increase difference in sales year would reduce the difference in prices. This result is counter-intuitive but explains a possible trend in falling normalized prices. The difference in FRM has an expected significant positive sign which implies increased differences in interest rates have negative effects on differences in prices. My linear models did not have expected sign on FRM which is now captured in the first two difference models. Interestingly, although increase in OE numbers in CFSD has positive and significant effects on the linear terms of both inner and outer buffer zones while negative signs on the quadratic terms; I will show later the numbers of OE which change the curvature for the two zones are different. Also the effects on differences in prices are positive and significant for TUSD, AUSD, TVUD and CFSD.

For Model VII, which includes the Charter Enrollments and captures observations from 2006 has counter-intuitive signs for both difference in years of sale and FRM. But I obtain almost significant and expected signs of the difference in OE numbers for both outer and inner buffer zone. Also note that CE interacted with inner zone gives a positive effect on the difference in house values.

Another interesting finding is that the difference models for the full dataset and the restricted dataset give identical parameter estimates whether we use prices or prices per square footage as an entry in the dependent variable. This is due to the fact that livable area hasn't changed over time and hence in the log-linear difference format gets washed away. Hence, we are essentially considering Model V and VII in the first difference approach.

6.2 Marginal Effects

In this subsection I only present the marginal effects of OE in the two buffer zones considered in this study. Other marginal effects are calculated similarly but are omitted. The method implemented to calculate the marginal effects is given by the simplistic example of a dependent variable $\log(y)$ which is a non-linear function of x:

$$\log y = \beta_0 + \beta_1 X \cdot D + \beta_2 X^2 \cdot D \tag{6.1}$$

Marginal effect on $\log y$:

$$\frac{\partial \log y}{\partial X} = \beta_1 D + 2\beta_2 X \cdot D; \text{ where } D = 1, \tag{6.2}$$

$$= \beta_1 + 2\beta_2 X \tag{6.3}$$

Marginal effect on y:

$$\frac{\partial y}{\partial X} = y\left[\beta_1 + 2\beta_2 X\right];\ \text{where}\, D = 1, \tag{6.4}$$

Critical value of x:

$$x_{critical} = -\frac{\beta_1}{2\beta_2}\ where\ \text{where}\, D = 1, \tag{6.5}$$

For this model, we are considering x to be OE in CFSD interacted with outer and inner buffer dummies. The calculation of marginal effects on difference in prices for change in open enrollment is difficult to evaluate as the log-difference model is a ratio of prices. The marginal effects are calculated at median prices for each zone in the time-framework considered while an average of OE in CFSD is considered which is rounded off to 489 students. Table 6.1 gives the median home values in the buffer zones:

buffer zone		All Observations	Obs from 2006
Outer-buffer	Median Sale Prices	\$209,110.60	\$222,186.5
	Median Sale Price per sq. ft. of livable area	\$97.96	\$106.29
Inner-buffer	Median Sale Prices	\$280,046.13	\$309,898.24
	Median Sale Price per sq. ft. of livable area	\$114.6	\$125.45

Table 6.1: Median Prices in buffer zones for both Models

Using eqn 6.3 we calculate the marginal effects of OE on log (prices) in both the critical zones considered in this paper, equation 6.4 gives the marginal effects on prices. While it is difficult to calculate the marginal effects of difference in OE on actual changes in prices, but we can certainly calculate the marginal effects on log of differences in prices. Fig 6.2 gives the marginal effects in the standard log-linear framework. While all houses in the buffer zone are in CFSD, houses in the outer buffer zone can either be in TUSD or AUSD. Hence, marginal effects on the outer zone also depends if the house is in TUSD or AUSD. In Model I, we can conclude that an increase in one more open enrolled student in CFSD would increase prices in outer buffer zone by \$29 if the house is in TUSD and \$61 if the house is in AUSD. We also see that in inner buffer zone prices increase by \$44 in Model I which is counter-intuitive because we expect prices to go down in the inner buffer zone. Similar results are observed for Model II where a price per square foot of living area is considered as a dependent variable. Model III which uses the dataset from 2006 gives a significant increase of \$47 for AUSD while a decrease of \$176 for TUSD (counter-intuitive). Also a decrease in prices is observed for the inner buffer zone by \$240. The additional variables of interest in the models with dataset from 2006 are ones for Charter school enrollments and its respective interactions. Once controlled for that, we find the expected sign for inner buffer zone but not for outer zone in TUSD. While some of the effects don't have expected signs, a general

pattern is that whatever the direction of change is, inner buffer zone is most affected negatively then outer zone in TUSD followed by outer zone in AUSD for positive changes in OE.

Effect of Open Enrollment Standard Models		log(Price)	Price	log (Price per sqaure feet of Livable Area)	Price per sqaure feet of Livable Area
Model I	Outer Buffer Zone if in TUSD	0.00014	29		
	Outer Buffer Zone if in AUSD	0.00029	61		
	Inner Buffer Zone	0.00016	44		
Model II	Outer Buffer Zone if in TUSD			0.00014	0.0144
	Outer Buffer Zone if in AUSD			0.00024	0.024
	Inner Buffer Zone			0.00017	0.021878
Model III	Outer Buffer Zone if in TUSD	-0.00079	-176.554		
	Outer Buffer Zone if in AUSD	0.000213	47.41238		
	Inner Buffer Zone	-0.00077	-240		
Model IV	Outer Buffer Zone if in TUSD			-0.00026	-0.0276
	Outer Buffer Zone if in AUSD			-0.00061	-0.06534
	Inner Buffer Zone			0	0

Figure 6.1: Marginal Effects of OE in buffer zones

Due to the non-linearity in the models it is also important to know the critical values of OE which would change the curvature of housing price change. Another simple equation which shows how I calculate the critical values of x is equation (6.5).

Equation (6.5) calculates the critical values of OE when the dependent variable changes its curvature. We have observed significant non-linear entries in Models I II for the outer buffer zone and none for the inner buffer zone in the log-linear models. For the differenced models we observed non-linearity in both outer and inner buffer zones for both Models V and VII.

In Model I we can observe that OE will increase prices in the outer buffer zone till about 1,177 if the house is in TUSD and 2,405 if it is in AUSD respectively and then it would decline the prices. It also shows that the critical number for inner zone would be 4391. Similarly, in Model III we obtain that prices would go up perpetually in the outer buffer zone and go down in the inner zone as the functional forms are linear. For the difference models, Models V the similar numbers for outer buffers are 1,477 if house is in TUSD and 2026 if in AUSD. Model VII gives 4144 and 2638 respectively for both these districts in the outer zone. While for the inner buffer zone in the difference models we can obtain that prices go down significantly after 1562

students in OE in CFSD from Model V and Model VII gives a critical value of 3915.

Chapter 7

Conclusions and Future Work

This paper is a modest attempt to contribute to the existing literature on housing prices capitalization in the U.S. due to changes in public policies in the education sector. Considering a twelve years dataset from 10 school districts of Pima County, Arizona this paper attempts to evaluate the effect of Open Enrollment in the best school district on housing prices capitalization in the neighborhood districts. In this paper I have addressed the issues usually considered in hedonic models of housing prices by considering the standard log-linear approach and a difference methodology. The problem encountered in most of the studies is removal of unobserved neighborhood characteristics which might bias the parameter estimates in a standard hedonic analysis. Using the panel nature of the dataset this study explores the difference approach to wash out the unobserved heterogeneity in houses which are time-invariant. Finally, I focused on those houses which are in and around the boundary of CFSD by creating an outer buffer zone which encompassed a 2 mile stretch outside its boundaries and a 1 mile stretch within the boundary. It also included removal of houses in the north-eastern boundary which is covered mostly by the Catalina Foothills for the inner buffer zone.

My results suggest that in the standard log-linear model we observe that housing prices have appreciated in the outer buffer zone significantly but OE might have a negative impact if the number exceeds a certain critical value. It also suggests that OE has insignificant and mildly negative impacts on houses which are within the inner boundary. The results varied slightly due to model variations but mostly the standard log-linear approach could capture some positive effect on the houses in the outer buffer zone. Also, I observed that appreciation in house values due to OE is higher in AUSD than in TUSD.

Since the log-linear model can be criticized on the grounds of omitted variables bias due to the inability to capture all the unobserved time invariant neighborhood characteristics, this paper shows the results of difference models of the standard log-linear models estimated. This second approach obtains concavity of the effect of OE in outer buffer zone which is consistent with the findings of the original models and concavity in effect of OE in the inner buffer zone

but with different critical values. Here I observe that prices will go down in the inner buffer zone but with more OE than in TUSD, which is counter-intuitive. But, the results are expected if we compare AUSD with TUSD. While it is difficult to choose the best model from the eight models estimated, a general overview from most of the models is that housing prices appreciated in the outer-buffer zone significantly but might decline after a certain critical enrollment; the inner-zone might have had a decline but it is not consistent with all the models. It is quite evident that the controlling for the charter school enrollments, we get different results.

First, the difference models might be preferred over the standard log-linear models because it captures certain unobserved heterogeneity by washing them out. If we believe the results from the difference models over the standard ones, including the charter enrollments provide different results. Second, it is also interesting to note that if OE has increased housing prices in the inner buffer zone but it decreases certainly after a threshold value. This indicates that OE might initially increase prices in both the zones due to other unobserved factors such as awareness, family composition etc. but it certainly decreases the prices in the inner buffer after a critical enrollment which is much lesser for the outer buffer in AUSD. The trend of OE in CFSD shows that the critical level of students has been reached while the critical level for the outer buffer in AUSD is yet to be reached. I also find counter-intuitive results for TUSD. The decline in the outer zone is difficult to explain but it might show that over time with expansion of charter schools and magnet schools, OE might affect negatively.

Future work remains to be done in terms of coming up with estimation techniques such that unobserved time varying characteristics can also be controlled for as the method implemented here can only remove time-invariant characteristics. One might thus incline towards this approach rather than the one used in the literature. Finally, to capture intra-district effects one might try matching schools with houses and see how distance from a desired school to the outer-buffer zone affects people's decision on paying a premium for the boundaries outside the best school districts.

References

Arizona Department of Education Research and Evaluation http://www.azed.gov/research-evaluation/

Arizona State Legislature: http://www.azleg.gov/FormatDocument.asp?inDoc=/ars/15/00816-01.htmTitle=15DocType=ARS

Bark, R.H., Osgood, D.E., Colby, B.G., Katz, G. and Stromberg, J., Habitat preservation and restoration: Do homebuyers have preference for quality habitat? , Ecological Economics 68 (2009) 1465-1475

Black, Sandra E., Do Better Schools Matter? Parental Valuation of Elementary Education, Quarterly Journal of Economics 114(2) (1999) 577-599.

Bogart, William T., and Brian A. Cromwell, How Much is a Neighborhood School Worth? , Journal of Urban Economics 47(2) (2000) 280-305.

Bourne, Kimberley L. , The effect of the Santa Cruz Riparian Corridor on Single Family home prices using the Hedonic Pricing Method, Master's Thesis submitted to Department of Agricultural and Resource Economics, University of Arizona (2007)

CFSD Open-Enrollment Numbers- courtesy Mary Kamerzell, Superintendant, Catalina Foothills School District

Downes, Thomas A. and Zabel, Jeffrey E., The impact of school characteristics on house prices: Chicago 1987-1991, Journal of Urban Economics 52 (2002) 1-225

Fossey, R. ,Open Enrollment in Massachusetts: Why families choose, Education Evaluation and Policy Analysis 16(3) (1994), 320-324

FRM data: http://www.hsh.com/

Funkhouser and Colopy, Minnesota's Open Enrollment Option, Educational Resources Information Center (1994)

Goldhaber, Dan D., School Choice: An Examination of the Empirical Evidence on Achievement, Parental Decision Making, and Equity, Educational Researcher 28(9) (1999), 16-25

Heckman, J.J., The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, Annals of Economics and Social Measurement 5(1976), 475-492

Kane,Thomas J. and Staiger, Douglas O., School Quality, Neighborhoods, and Housing Prices, American Law and Economics Review (2006) 183-212

Oates, W.E., The effects of property taxes and local public spending on property values: An

empirical study of tax capitalization and the Tiebout hypothesis, Journal of Political Economy 77 (1969) 957-971.

Pima County School District Map - "Figure 3.1" courtesy "The Pima County School Superintendent": http://www.schools.pima.gov/site/schools

Pima County GIS: http://gis.pima.gov/

Pima County Assessor's Office: http://www.asr.pima.gov/

Reback, Rendall, House prices and the provision of local public services: capitalization under school choice programs, Journal of Urban Economics 57 (2005) 275-301

Rosen, S., Hedonic prices and implicit markets: Product differentiation in pure competition, Journal of Political Economy 82 (1974) 34-55.

Rubenstein, Michael C., Minnesota's Open Enrollment Option, Educational Resources Information Center (1992)

S & P/Case-Schiller Home Price Indices: http://www.standardandpoors.com/indices/sp-case-shiller-home-price-indices/en/us/?indexId=spusa-cashpidff–p-us—-

Smith, Angela G. , Public School Choice and Open Enrollment: Implications for Education, Desegregation, and Equity, Journal of Law and Education 24(147) (1995)

The Pima County School Superintendent: http://www.schools.pima.gov/

Weimer,David L. and Wolkoff, Michael J., School Performance and Housing Values: Using Non-contiguous District and Incorporation Boundaries to identify School Effects, National Tax Journal 54(3) (2001) 231-253

Appendix

FULL MODEL (ALL OBSERVATIONS)				
Dependent Variable	Model I		Model II	
	Log(Price)		Log(Price/Livable Area)	
Variable	**Estimate**	**Pr > |t|**	**Estimate**	**Pr > |t|**
Intercept	9.50271(0.20201)	<.0001	3.13818(0.22461)	<.0001
d_new	0.03988(0.00173)	<.0001	0.03934(0.00172)	<.0001
ROOMS	-0.03733(0.00098996)	<.0001	-0.05854(0.00108)	<.0001
SQFT	0.00043233(0.00000359)	<.0001	-0.00000939(0.00000334)	0.0049
LANDSQFT	0.00000147(0.0000001238727)	<.0001	0.00000166(0.0000001325683)	<.0001
BATHFIXTUR	0.02216(0.00056102)	<.0001	0.02301(0.00053205)	<.0001
d_pool	0.10336(0.00185)	<.0001	0.09456(0.00179)	<.0001
d_patio	0.08447(0.002)	<.0001	0.0656(0.00202)	<.0001
d_garage	0.11901(0.00251)	<.0001	0.09928(0.0025)	<.0001
d_ev_cool	0.46321(0.20015)	0.0207	0.43892(0.22284)	0.0489
d_ac_cool	0.59009(0.20014)	0.0032	0.53364(0.22282)	0.0166
d_fair_quality	0.24278(0.01065)	<.0001	0.1416(0.01043)	<.0001
d_good_quality	0.26607(0.01077)	<.0001	0.16047(0.01053)	<.0001
d_excellent_quality	0.29133(0.01874)	<.0001	0.3083(0.01743)	<.0001
FRM	0.02302(0.00289)	<.0001	0.02439(0.00288)	<.0001
d_tucson	0.39266(0.03088)	<.0001	0.42612(0.03002)	<.0001
d_marana	0.38385(0.0309)	<.0001	0.40288(0.03004)	<.0001
d_flowingwells	0.35121(0.03113)	<.0001	0.38358(0.03031)	<.0001
d_amphitheather	0.45946(0.03092)	<.0001	0.48291(0.03005)	<.0001
d_sunnyside	0.16233(0.031)	<.0001	0.20704(0.03016)	<.0001
d_tanque	0.57302(0.03122)	<.0001	0.60167(0.02994)	<.0001
d_cat	0.7317(0.03188)	<.0001	0.7877(0.03088)	<.0001
d_vail	0.35642(0.03095)	<.0001	0.37191(0.0301)	<.0001
d_sahuarita	0.24727(0.03099)	<.0001	0.27697(0.03015)	<.0001
d_outer_buffer	0.13624(0.00854)	<.0001	0.13444(0.00856)	<.0001
d_inner_buffer	-0.09662(0.01627)	<.0001	-0.11639(0.01402)	<.0001
d_outer_buffer * Cat_Op_Enroll	0.0002063(0.00004869)	<.0001	0.00021977(0.00004894)	<.0001
d_inner_buffer * Cat_Op_Enroll	0.00004397(0.00008429)	0.6019	0.0000583(0.00007061)	0.409

Variable	Estimate	Pr > |t|	Estimate	Pr > |t|
(d_outer_buffer * Cat_Op_Enroll)^2	-0.0000000619001(0.00000003521847)	0.0788	-0.0000000680579(0.0000000354374)	0.0548
(d_inner_buffer * Cat_Op_Enroll)^2	-0.0000000178971(0.0000000584202)	0.7593	-0.0000000346907(0.00000004883915)	0.4775
d_tanque* Cat_Op_Enroll	0.00007899(0.00001759)	<.0001	0.00007915(0.00001527)	<.0001
d_tucson* Cat_Op_Enroll	-0.00006064(0.00000465)	<.0001	-0.00006365(0.00000463)	<.0001
d_cat* Cat_Op_Enroll	0.00015717(0.000017)	<.0001	0.0001744(0.00001573)	<.0001
d_amphitheather* Cat_Op_Enroll	0.00009143(0.00000561)	<.0001	0.00009219(0.00000549)	<.0001
d_2002	0.03124(0.00309)	<.0001	0.03264(0.00307)	<.0001
d_2003	0.06013(0.00437)	<.0001	0.06101(0.00435)	<.0001
d_2004	0.0577(0.00426)	<.0001	0.05979(0.00424)	<.0001
d_2005	-0.03172(0.0043)	<.0001	-0.02842(0.00428)	<.0001
d_2006	-0.13457(0.00319)	<.0001	-0.1302(0.00319)	<.0001
d_2007	-0.12526(0.00342)	<.0001	-0.12031(0.00344)	<.0001
d_2008	0.05729(0.00383)	<.0001	0.05994(0.00386)	<.0001
d_2009	0.28757(0.00881)	<.0001	0.2927(0.00878)	<.0001
d_2010	0.16251(0.00751)	<.0001	0.16795(0.00748)	<.0001
d_2011	0.10656(0.00859)	<.0001	0.11125(0.00854)	<.0001
d_2012	0.07555(0.01083)	<.0001	0.07818(0.01078)	<.0001
F-value	9768.00		2243.40	
R-squared	0.7163		0.3670	
Adj R-squared	0.7162		0.3668	
Observations	170291		170291	

Figures in parenthesis correspond to heteroskedasticity consistent standard errors

MODEL WITH CHARTER SCHOOL DATA (OBSERVATIONS FROM 2006)				
Dependent Variable	Model III		Model IV	
	Log(Price)		Log(Price/Livable Area)	
Variable	**Estimate**	**Pr > |t|**	**Estimate**	**Pr > |t|**
Intercept	9.01457(0.2307)	<.0001	2.64945(0.24214)	<.0001
d_new	0.0179(0.00293)	<.0001	0.01655(0.00298)	<.0001
ROOMS	0.00151(0.00011412)	<.0001	0.0013(0.00010972)	<.0001
SQFT	0.00037175(0.00000396)	<.0001	-0.00009689(0.00000402)	<.0001
LANDSQFT	0.00000171(0.0000001823236)	<.0001	0.00000202(0.0000002079231)	<.0001
BATHFIXTUR	0.01978(0.00086411)	<.0001	0.01903(0.00085598)	<.0001
d_pool	0.11458(0.00312)	<.0001	0.10646(0.00311)	<.0001
d_patio	0.09087(0.00353)	<.0001	0.06947(0.00357)	<.0001
d_garage	0.12971(0.00412)	<.0001	0.10462(0.00411)	<.0001
d_ev_cool	0.1756(0.22659)	0.4384	0.08528(0.2384)	0.7205
d_ac_cool	0.34379(0.22658)	0.1292	0.22038(0.23838)	0.3552
d_fair_quality	0.27636(0.01656)	<.0001	0.16768(0.01653)	<.0001
d_good_quality	0.29043(0.01673)	<.0001	0.17435(0.01666)	<.0001
d_excellent_quality	0.37826(0.03021)	<.0001	0.40147(0.02972)	<.0001
FRM	0.08183(0.00534)	<.0001	0.08479(0.00537)	<.0001
d_tucson	-1.6412(0.14495)	<.0001	-1.60983(0.14661)	<.0001
d_marana	0.43994(0.04714)	<.0001	0.45249(0.04657)	<.0001
d_flowingwells	0.39362(0.04777)	<.0001	0.414(0.04727)	<.0001
d_amphitheather	-0.58304(0.1657)	0.0004	-0.56948(0.16743)	0.0007
d_sunnyside	0.16743(0.04732)	0.0004	0.19984(0.0468)	<.0001
d_tanque	0.26828(0.55284)	0.6275	-0.05508(0.48)	0.9086
d_cat	-0.10618(0.43303)	0.8063	-0.02824(0.42004)	0.9464
d_vail	0.42907(0.04731)	<.0001	0.43662(0.04676)	<.0001
d_sahuarita	0.25442(0.04733)	<.0001	0.26901(0.04681)	<.0001
d_outer_buffer	0.23951(0.35318)	0.4977	-0.01322(0.3695)	0.9715
d_inner_buffer	-1.13694(0.68078)	0.0949	-0.45987(0.61021)	0.4511
outer_eff_Op	0.00029538(0.00014216)	0.0377	0.00020791(0.00014836)	0.1611
inner_eff_Op	-0.00046877(0.00028307)	0.0977	-0.00014075(0.00026319)	0.5928
d_outer_buffer * Cat_Op_Enroll	-0.0000000800006(0.0000000529332)	0.1307	-0.000000101674(0.00000005442398)	0.0617

Variable	Estimate	Pr > |t|	Estimate	Pr > |t|
d_inner_buffer * Cat_Op_Enroll	-0.0000000266389(0.00000009379542)	0.7764	-0.0000000153085(0.00000007676806)	0.8419
(d_outer_buffer * Cat_Op_Enroll)^2	-0.00008199(0.00027108)	0.7623	-0.00025486(0.00023451)	0.2771
(d_inner_buffer * Cat_Op_Enroll)^2	-0.00109(0.0000662)	<.0001	-0.00109(0.00006723)	<.0001
d_tanque* Cat_Op_Enroll	-0.00030406(0.00020529)	0.1386	-0.00028432(0.0002001)	0.1554
d_tucson* Cat_Op_Enroll	-0.00046446(0.00007636)	<.0001	-0.00046403(0.00007744)	<.0001
d_outer_buffer * Charter_Enroll	-0.00000131(0.00000402)	0.7442	0.00000166(0.00000421)	0.6932
d_inner_buffer * Charter_Enroll	0.00001233(0.00000774)	0.1109	0.00000406(0.00000698)	0.5611
d_Amphitheather * Charter_Enroll	0.00001296(0.00000185)	<.0001	0.00001298(0.00000187)	<.0001
d_Tanque * Charter_Enroll	0.00000392(0.00000648)	0.545	0.00000795(0.00000562)	0.1569
d_Tucson * Charter_Enroll	0.00002464(0.0000016)	<.0001	0.00002454(0.00000162)	<.0001
d_Cat * Charter_Enroll	0.00001063(0.00000502)	0.0342	0.00001042(0.00000488)	0.0326
d_2008	0.17803(0.00341)	<.0001	0.17558(0.00351)	<.0001
d_2009	0.48847(0.0095)	<.0001	0.49102(0.00954)	<.0001
d_2010	0.40599(0.00933)	<.0001	0.40969(0.00941)	<.0001
d_2011	0.33764(0.01074)	<.0001	0.34017(0.0108)	<.0001
d_2012	0.33128(0.01453)	<.0001	0.3331(0.01462)	<.0001
F-value	3759.58		1005.74	
R-squared	0.6781		0.3605	
Adj R-squared	0.678		0.3601	
Observations	80343		80343	

Figures in parenthesis correspond to heteroskedasticity consistent standard errors

The model presented in the next page is an appended model which captures an additional effect of Open Enrollment in CFSD at outer buffer houses only in TUSD. Since the results are not much different, the interpretations remain similar to those from the ones already discussed.

Dependent Variable	Log(Price)		
Variable	**Estimate**	**t-value**	**Pr >|t|**
Intercept	9.50718	47.04	<.0001
d_new	0.04161	24.01	<.0001
ROOMS	-0.03721	-37.7	<.0001
SQFT	0.000432	120.51	<.0001
LANDSQFT	1.47E-06	11.86	<.0001
BATHFIXTUR	0.02185	39.01	<.0001
d_pool	0.10196	55.31	<.0001
d_patio	0.08363	41.86	<.0001
d_garage	0.11588	46.14	<.0001
d_ev_cool	0.46322	2.31	0.0207
d_ac_cool	0.58911	2.94	0.0033
d_fair_quality	0.24275	22.83	<.0001
d_good_quality	0.26586	24.74	<.0001
d_excellent_quality	0.29598	15.86	<.0001
FRM	0.0231	8.02	<.0001
d_tucson	0.39318	12.72	<.0001
d_marana	0.38577	12.48	<.0001

d_flowingwells	0.35089	11.26	<.0001
d_amphitheather	0.46105	14.9	<.0001
d_sunnyside	0.16189	5.22	<.0001
d_tanque	0.5773	18.48	<.0001
d_cat	0.73589	23.07	<.0001
d_vail	0.35844	11.57	<.0001
d_sahuarita	0.24867	8.02	<.0001
d_outer_buffer	0.13791	16.88	<.0001
d_inner_buffer	-0.0964	-5.94	<.0001
d_outer_buffer * Cat_Op_Enroll	1.06E-05	0.23	0.8216
d_inner_buffer * Cat_Op_Enroll	4.45E-05	0.53	0.5963
(d_outer_buffer * Cat_Op_Enroll)^2	-6.12E-08	-1.85	0.0642
(d_inner_buffer * Cat_Op_Enroll)^2	-1.86E-08	-0.32	0.7495
d_tucson*d_outer_buffer*Cat_Op_Enroll	0.000344	23.61	<.0001
d_tanque* Cat_Op_Enroll	7.85E-05	4.47	<.0001
d_tucson* Cat_Op_Enroll	-7E-05	-14.88	<.0001
d_cat* Cat_Op_Enroll	0.000157	9.26	<.0001
d_amphitheather* Cat_Op_Enroll	0.000119	21.06	<.0001
d_2002	0.03129	10.15	<.0001
d_2003	0.0602	13.82	<.0001

d_2004	0.05793	13.64	<.0001
d_2005	-0.03139	-7.32	<.0001
d_2006	-0.13413	-42.12	<.0001
d_2007	-0.12496	-36.56	<.0001
d_2008	0.05766	15.07	<.0001
d_2009	0.28811	32.74	<.0001
d_2010	0.16301	21.74	<.0001
d_2011	0.10788	12.59	<.0001
d_2012	0.07649	7.07	<.0001
F-value	9609.04		
R-squared	0.7175		
Adj R-squared	0.7174		
Observations	170291		

Difference Model with All Observations (Model V)

Dependent Variable	**Model V**	
	Log(Price)	
Variable	**Estimate**	**Pr > |t|**
dif_cat_yr	-0.02632(0.00164)	<.0001
dif_FRM	-0.01671(0.00293)	<.0001
dif_(d_outer_buffer*Cat_Op_Enroll)	0.00027007(0.00005147)	<.0001
dif_(d_inner_buffer*Cat_Op_Enroll)	0.0002157(0.00008347)	0.0098
dif_(d_outer_buffer*Cat_Op_Enroll)^2	-0.000000101673(0.00000003759459)	0.0068
dif_(d_inner_buffer*Cat_Op_Enroll)^2	-0.000000178431(0.0000000591559)	0.0026
dif_(d_Tucson*Cat_Op_Enroll)	0.00003029(0.00000545)	<.0001
dif_(d_Cat*Cat_Op_Enroll)	0.00028737(0.00001544)	<.0001
dif_(d_Tanque*Cat_Op_Enroll)	0.00016282(0.00002008)	<.0001
dif_(d_Amphitheather*Cat_Op_Enroll)	0.00014187(0.00000617)	<.0001
d_2002	0.14091(0.01351)	<.0001
d_2003	0.11517(0.00675)	<.0001
d_2004	0.09407(0.00517)	<.0001
d_2005	0.0319(0.0046)	<.0001
d_2006	-0.00767(0.00646)	0.2351
d_2007	0.02145(0.00781)	0.0061
d_2008	0.18525(0.00866)	<.0001
d_2009	0.31308(0.01034)	<.0001
d_2010	0.21945(0.01336)	<.0001
d_2011	0.17062(0.01302)	<.0001
d_2012	0.13711(0.0142)	<.0001
F-value	253.00	
R-squared	0.1143	
Adj R-squared	0.1138	
Observations	41254	

Figures in parenthesis correspond to heteroskedasticity consistent standard errors.
As clarified, Model VI estimates are identical to Model V and hence are not reported.

Difference Model with observations from 2006 (Model VII)

Dependent Variable	**Model VII**	
	Log(Price)	
dif_cat_yr	-0.0048(-0.00506)	0.343
dif_FRM	0.12206(-0.01014)	<.0001
dif_(d_outer_buffer*Cat_Op_Enroll)	0.00064516(0.00026746)	0.0159
dif_(d_inner_buffer*Cat_Op_Enroll)	-0.00072173(0.00045272)	0.1109
dif_(d_outer_buffer*Cat_Op_Enroll)^2	-0.000000195873(0.00000009043613)	0.0303
dif_(d_inner_buffer*Cat_Op_Enroll)^2	-0.000000209206(0.0000001192419)	0.0794
dif_(d_Tucson*Cat_Op_Enroll)	0.0003881(0.00011958)	0.0012
dif_(d_Cat*Cat_Op_Enroll)	0.00236(0.00036152)	<.0001
dif_(d_Tanque*Cat_Op_Enroll)	0.0017(0.00046051)	0.0002
dif_(d_Amphitheather*Cat_Op_Enroll)	0.00097805(0.00013759)	<.0001
dif_(d_inner_buffer*Charter_Enroll)	0.00002409(0.00001179)	0.0411
dif_(d_outer_buffer*Charter_Enroll)	-0.00000513(0.00000754)	0.4963
dif_(d_Amphitheather*Charter_Enroll)	-0.0000201(0.00000334)	<.0001
dif_(d_Tanque*Charter_Enroll)	-0.00003556(0.00001102)	0.0013
dif_(d_Tucson*Charter_Enroll)	-0.00000872(0.00000289)	0.0026
dif_(d_Cat*Charter_Enroll)	-0.00005129(0.00000866)	<.0001
d_2007	0.20468(0.01225)	<.0001
d_2008	0.20057(0.01265)	<.0001
d_2009	0.44974(0.01485)	<.0001
d_2010	0.30459(0.02202)	<.0001
d_2011	0.31804(0.01906)	<.0001
d_2012	0.35857(0.01993)	<.0001
F-value	126.17	
R-squared	0.1729	
Adj R-squared	0.1715	
Observations	13303	

Figures in parenthesis correspond to heteroskedasticity consistent standard errors.
As clarified, Model VIII estimates are identical to Model VII and hence are not reported.

Since we are considering a difference model, it is important to look at the summary statistics of key variables for the different categories of single-houses sold: Sold Once, Sold Twice, Sold Thrice, and Sold more than three times. The following tables provide the summary statistics of these separate categories.

Table: Houses Sold Once

Variable	N	Mean	Median	Minimum	Maximum	Std Dev
ROOMS	69,873	6.9	7	1	50	1.58
SQFT	69,873	1,868.5	1727	224	9,949	689.5
LANDSQFT	69,873	14,887.2	7755	140	1,718,375	29,651.8
BATHFIXTUR	69,873	7.3	7	3	30	2.43
Price	69,873	160,451.9	129,205.6	6,624.71	7,643,799	135,306
Price_liv_area	69,873	82.57	76.47	4.1	4,065.9	44.5

Table: Houses Sold Twice

Variable	N	Mean	Median	Minimum	Maximum	Std Dev
ROOMS	64,234	6.9	7	1	19	1.55
SQFT	64,234	1,841.9	1,707	306	10,901	652.9
LANDSQFT	64,234	12,693.4	7,526	1,152	2,006,163	257240.7
BATHFIXTUR	64,234	7.28	7	3	28	2.36
Price	64,234	166,396.4	124,873.8	4945.1	7,643,799	343,695.5
Price_liv_area	64,234	88.35	74.68	3.6	6306.7	199

Table: Houses Sold Thrice

Variable	N	Mean	Median	Minimum	Maximum	Std Dev
ROOMS	27,411	6.8	7	1	19	1.53
SQFT	27,411	1,766.9	1,628	434	7977	617.7
LANDSQFT	27,411	11,383.4	7,470	1,204	1,204,487	20944
BATHFIXTUR	27,411	7.05	7	3	22	2.29
Price	27,411	143,165.8	118,310.5	4,532.5	7,643,799	137,397.9
Price_liv_area	27,411	78.77	73.83	2.4	5,168.2	70.34

Table: Houses sold more than Thrice

Variable	N	Mean	Median	Minimum	Maximum	Std Dev
ROOMS	8,773	6.6	6	1	15	1.51
SQFT	8,773	1,698.8	1,560	576	6,345	605.37
LANDSQFT	8,773	10,683.8	7,517	1,768	216,245	13,568.6
BATHFIXTUR	8,773	6.71	6	3	27	2.36
Price	8,773	131,667.9	108,842.6	9,356.3	956,604.9	86,413.7
Price_liv_area	8,773	74.96	71.66	4.9	292.6	28.25

Essay II: Sub-sample Estimators of Big Data

Contents

List of Figures

Abstract

Last two decades have witnessed huge datasets in the fields of research, business and finance with thousands of variables and billions of records. This presents difficulty in standard regression analysis as the analysts face costs in terms of time constraint and computer memory size. In this paper, I propose estimators using two procedures to estimate datasets of three different sizes when the dependent variable is binary in nature. Both theoretically and empirically, this paper shows how estimation time can be significantly reduced using the proposed techniques resulting in linear unbiased estimators which have variances not much different from the classic model. While the literature relied on splitting the entire dataset in blocks, I use replicates in form of samples from the dataset using simple random sampling. Finally this paper endorses one procedure above other on grounds of efficiency.

Chapter 1

Introduction

With the revolution in information technology, the last two decades have witnessed routine collection of systematically generated data in fields of research, business and finance. Nowadays, databases are attributed with hundreds of variables, billions of records and terabytes of information. Information is increasing and more than 2.5 billion data are created everyday [Wikipedia]. In Information Technology the terminology of "BIG DATA" is used to designate those kinds of datasets which grows so large that it becomes extremely difficult to capture, analyze and share them. Barclaycard (UK) has more than 350 million transactions a year, Wal-Mart makes over 7 billion transactions a year, and AT&T carries over 70 billion long distance calls annually [Hand et al.]. Scientists regularly face these problems in genomics, meteorology, complex physics, simulations, biological research and big corporate giants are spending huge sums to come up with techniques and technology to handle such datasets. Oracle, IBM, Microsoft, SAP spent more than $15 billion on data management and analytics [Economist, 2010]. The immediate consequence of such databases is difficulties in storing, visualizing and analyzing, using classical data analysis methods. The prime reason is limitation of computer memory and computational time which hinders the ability to utilize the entire dataset.

The primary tasks in analyzing large datasets include data processing, classification, summarization, visualization, association, correlation and regression. There is a rich literature on how huge datasets can be aptly handled with little or no apologies in standard regression analysis. Since regression analysis is not straightforward for massive datasets, the literature shows both empirically and theoretically how splitting the datasets into appropriate blocks can be useful in minimizing the computational time but still resulting optimal estimation results. The statistical perspectives in analyzing massive datasets via Bayesian approach include works by Elder Pregibon, Glymour et al., Ridgeway and Madigan, Jackman, Balakrishnan and Madigan. The econometric approach can be found in the works of Fan et al. and Li et al. Literatures in the econometric analysis include standard models using OLS procedure. For example, Fan et al. shows using standard OLS estimation how a huge dataset can be split into several mutually

exclusive blocks and then estimates from each block be combined to find out final parameter estimates. They in particular combine the estimates by minimizing the variance of the final estimator.

In this paper I use binary dependent variable model instead of standard continuous dependent variable and use datasets of different sizes to study how we can handle them. Consequently, my study differs from the previous work as it now needs to be treated as a linear probabilty model as opposed to standard continuous variable model with OLS technique. I randomly generate data of various sizes- 100 thousand observations, 1 million observations and 10 million observations and use the model to come up with the parameter estimates. In all these models of different sizes, I first regress using the entire data and then do the same by segregated portions of the entire dataset. The portions of dataset chosen are different from the ones used in the literature. In this study, replicates were formed by randomly choosing data from the existing dataset using simple random sampling as opposed to just merely splitting them in parts. This paper is divided into five sections. Section 2 describes how the data was generated and stored, section 3 focusing on the methodology used, section 4 summarizing the results and final section concludes.

Chapter 2

Data

Since the entire analysis and motivation for this kind of a study is pivoted crucially on the computer used for the analysis, its specification is very important as results would vary from one system to another. The computer used in this study is a DELL personal laptop with Intel (R) Core(TM) 2 Duo CPU, a 3.00 GB DDR2 RAM and Windows Vista Home Premium as it OS. All the regressions were run in the same computer to avoid any computer specific errors with the same specification.

The dataset is randomly generated using SAS 9.2 (32) software and stored in SAS library for all uses. The data includes four independent variables x_1, x_2, x_3, x_4 and an error term u which were generated as normal random variables, v as standard normal to fit into the standard assumptions of a probit model: $x_1 \sim N(0.567, 1.577536)$, $x_2 \sim N(1.936, 0.06589489)$,
$x_3 \sim N(-2.986, 8.392609)$, $x_4 \sim N(0.256, 0.183184)$,$v \sim N(0, 1)$

The model is as follows:

$$y_i = -2 + 0.8x_{i1} + 3.6x_{i2} + 1.3x_{i3} - 0.9x_{i4} + u_i \text{ where } i = 1(1)N \tag{2.1}$$

$$y_i = \begin{cases} 0, & \text{if } y_i^* \leq 0 \\ 1, & \text{if } y_i^* > 0 \end{cases} \tag{2.2}$$

where y_i is a latent variable.

$$y_i^* = \beta_i' x_i + v_i \text{where } v_i \sim N(0, 1) \tag{2.3}$$

First, data were generated for all the independent variables and the error term v. Using all these the data I generated the latent variable and finally using the latent variable, data on the binary variable y was generated using the above relationship between y and y^*. Three different datasets of sizes 100 thousand, one million and ten million were generated separately and analyzed. The set up of the model clearly shows that we should use probit model as the

errors are normally distributed with mean 0 and equal variance of 1. All the usual assumptions of the probit model have been satisfied and the theory is the standard one in the context of a probit model. Given that the errors are normally distributed, we use the likelihood function of the error coming from a standard normal distribution and maximize the log-likelihood to arrive at the parameter estimates.

Chapter 3

Methodology

This simple set up has been used to exploit how randomly selected replicates can be used to arrive at parameter estimates which are then compared to the actual estimates found from the results using the full dataset. There can three proposed approaches if the dataset cannot be analyzed using all the observations. As opposed to the ones available in the literature where dataset has been segregated into parts using mutually exclusive and exhaustive blocks, I have randomly selected samples from the dataset which has been replicated to form several samples of equal size. The sampling technique used is simple random sampling with replacement. So, for example, if 'N' is the population size, 'n' is the sample size, we can choose 'R' such samples. These are the three approaches that can be utilized:

1. Create mutually exclusive and exhaustive blocks from the population such that all the observations are utilized. Such kind of segregation has been adopted by Fan et al. and Li et al. in standard linear regression models.

2. Samples of size 'n', where $n \leq N$, can be chosen 'R' times such that $nR = N$. This is the approach that has been exploited in this paper.

3. Samples can be chosen such that $nR < N$. The argument against such a technique would be that if the computer memory is not constrained, one might want to implement the second approach and use replicates to exhaust N.

Once the parameter estimates are available from all these samples/blocks, the important concern would be to aggregate them and obtain final estimators for all the parameters. The aggregation at this stage can be done in two possible ways:

- **Procedure 1 (P1)**: The aggregate estimate for each parameter would be the just a simple average of all the estimates for that parameter as obtained from the samples. I call these estimators as the "simple mean estimators".

- **Procedure 2 (P2)**: The aggregate estimate in this second procedure is obtained by taking a weighted average of the estimates of all the samples. These can be called as the "variance minimized estimators".

I will first theoretically represent the above two proposed procedures and briefly discuss their pros and cons in terms of their appropriateness in this context.

Simple Mean Estimators

This procedure is fairly easy to comprehend and carry out but it has some major flaws in terms of interpretation. The proposition in this procedure is that the final parameter estimates can be obtained by just taking the simple arithmetic mean of all estimates from the replicates for each parameter. Hence, if $\widehat{\beta}_i^1, \widehat{\beta}_i^2, \ldots, \widehat{\beta}_i^p$ are the estimates obtained for parameter β_i from each replicate j, where $j = 1(1)p$ then the simple means estimator would be given by,

$$\overline{\widehat{\beta}_i} = \frac{\Sigma_{j=1}^{p} \widehat{\beta}_j}{p} \tag{3.1}$$

where $j = 1(1)p$, $i = 1(1)k$.

Clearly, this final estimator is linear and unbiased. The variances of $\overline{\widehat{\beta}_l}$ can be obtained by taking the simple average of variances obtained from each replicates, i.e.,

$$V\left(\overline{\widehat{\beta}_i}\right) = \frac{\Sigma_{j=1}^{p} V\left(\widehat{\beta}_j\right)}{p} \tag{3.2}$$

where $j = 1(1)p$, $i = 1(1)k$.

and hence the variance obtained will not be a true representative of the actual variance of the simple mean estimator. But since this procedure is fairly easy to carry out, we shall the results obtained from this procedure to the results from variance minimized estimators as well as the results obtained after using the entire observations.

Variance Minimized Estimators

In this procedure, the final estimate is a weighted average of the estimates from each replicate, the weights being found by minimizing the variance of the final estimate. Also, we want the final variance minimized estimate, $\widetilde{\beta}_l$, to be linear and unbiased. The weights for each replicate are hence obtained:

$$Minimize\, Var(\widetilde{\beta}_i) = \sum_{j=1}^{p} w_j^2 Var(\widehat{\beta}_j^i) \tag{3.3}$$

where $j = 1(1)p$, $i = 1(1)k$.

subject to

$$\sum_{j=1}^{p} w_j = 1 \tag{3.4}$$

We need the constraint to make the estimator unbiased. The weights for each replicate of a parameter is given as,

$$w_{j,i} = \frac{\prod_{j=1}^{p} Var(\widehat{\beta}_j^i)}{Var(\widehat{\beta}_j^i) \sum_{h=1}^{p} \left(\prod_{-h} Var(\widehat{\beta}_j^i) \right)} \tag{3.5}$$

where $j, h = 1(1)p$, $i = 1(1)k$.

Proof: To find out the final weights for each replicate, the Lagrangian function is first set up,

$$\mathcal{L} = \sum_{j=1}^{p} w_j^2 Var(\widehat{\beta}_j^i) + \lambda \left[1 - \sum_{j=1}^{p} w_j \right] \tag{3.6}$$

The first-order necessary conditions are given as,

$$\frac{\partial \mathcal{L}}{\partial w_j} = 2 w_j Var(\widehat{\beta}_j^i) - \lambda = 0 \tag{3.7}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \left[1 - \sum_{j=1}^{p} w_j \right] = 0 \tag{3.8}$$

By solving these $(p+1)$ equations, (3.5) is obtained. The second order sufficient conditions are very easy to verify and hence has not been shown in this paper. This completes the proof. □

Hence, the resulting estimates are $\widetilde{\beta}_i = \sum_{j=1}^{p} w_j \widehat{\beta}_j^i$ with the final variances being equal to $Var(\widetilde{\beta}_i) = \sum_{j=1}^{p} w_j^2 Var(\widehat{\beta}_j^i)$, where the weights are obtained from the exercise above. It is easy to note that the variance minimized estimator is not only linear but also unbiased by virtue of the fact that all weights add up to 1.

We will compare the results obtained from this procedure to the first simple means procedure as wells as the results from all the observations.

Chapter 4

Results

In this section I discuss and compare the results found out from the two procedures mentioned above and finally compare them with the estimates found from the regressions using the entire dataset. There are two important aspects that motivate this study: first, if the proposed estimators are good enough instead of the ones found from the regression of the entire dataset in terms of their estimated values and the variances; second, if the proposed procedures are fairly less time consuming than the classical technique used in regression analysis, i.e., using all observations available. The time components of the results are important because if computer memory is constrained such that it cannot perform the regression using all observations, the other proposed procedures can come handy due to reasons stated later.

This study comprises of analysis using randomly generated datasets on the probit model discussed above with three different sets of observations: 100000 (one hundred thousand), 1000000 (one million) and 10000000 (ten million). In procedures 1(simple means) & 2 (variance minimized weighted average), I have varied the number of replications (R) and also the number of observations in each replication (n) such that N = nR for each dataset. Fig 4.1 shows the time taken by each of the two procedures for 'one hundred thousand' observations along with the probit regression using all observations. I report both the "Real time" and "CPU time" taken by the computer to run the programs where the first denotes the elapsed time, i.e., the time taken by 'wall clock' and the later denotes the time taken by the processor to execute the written code respectively (Fullstimer SAS option).

In fig 4.2 and fig. 4.3, similar results for datasets of 'one million' and 'ten million' are shown respectively.

It is quite evident from the results presented above that both the proposed procedures take longer time (except once) to execute the program when N = 100000 (Fig 4.1) no matter what the replication number is chosen. But, time taken by both the procedures is significantly lesser than the classical approach if N gets larger and if we choose n and R appropriately. So, for example, time elapsed in executing the regression using 10 million observations is 16 minutes

N = 100000						
	All observations		Procedure 1		Procedure 2	
Sample Size (n) & Replicates (R)	Real time	CPU time	Real time	CPU time	Real time	CPU time
N = 100000	2.86	2.85				
n = 100000, R = 1			3.07	2.63	2.84	2.71
n = 50000, R = 2			3.19	2.97	3.5	3.13
n = 10000, R = 10			3.89	2.82	4.32	2.87

Figure 4.1: Time taken by different procedures for 'One hundred thousand' observations

N = 1000000 (1 million)						
	All observations		Procedure 1		Procedure 2	
Sample Size (n) & Replicates (R)	Real time	CPU time	Real time	CPU time	Real time	CPU time
N = 1000000	36.49	32.54				
n = 1000000, R =1			32.87	31.87	31.33	30.53
n = 500000, R=2			32.22	30.09	30.03	29.67
n = 100000, R =10			31.71	28.68	31.09	28
n = 50000, R = 20			34.79	28.68	37.5	28.81
n = 10000, R =100			45.39	29.49	40.9	27.62

Figure 4.2: Time taken by different procedures for 'One million' observations

and 3.6 seconds while Procedure 1 and Procedure 2 takes 5 minutes 26.4 seconds and 5 minutes 6.7 seconds respectively when number of replications is 20 with each having half million observations. Thus, the gains in terms of time become evident from the procedures if the dataset is larger than usual. From Fig 4.1, we see that if dataset is as large as one hundred thousand observations, the proposed procedures fail in terms of saving the costs of longer time. As the dataset size goes larger, there is considerable time difference between the proposed estimators and the estimators using all observations but it crucially depends also on the number of replications chosen and observations in each replication. The following graphs show how the elapsed time varies with the 'R' and 'n':

It is quite evident from the above plots that time depends on the replication numbers and also on the number of observations in both P1 and P2. When the dataset is large enough, in this case of 1 million and 10 million observations, as R is increased time taken decreases, attains a minimum and then again increases. Also, the time taken is minimized by P2 despite the fact that P1 is fairly easier to carry out as a process. This feature is not quite obvious from the first dataset as both P1 and P2 are claimed to have failed in terms of minimizing time due to its smaller size. With large datasets, it may be hypothesized that time would exhibit almost a U-shaped curve with R. So in order to minimize time with any of the proposed procedures,

N = 10000000 (10 million)						
	All observations		Procedure 1		Procedure 2	
Sample Size (n) & Replicates (R)	Real time	CPU time	Real time	CPU time	Real time	CPU time
N = 10000000 (10 million)	16:03.6	15:51.4				
n = 10000000, R =1			16:13.9	15:44.8	16:17.6	15:41.0
n = 5000000, R = 2			20:07.9	19:19.3	17:03.5	17:05.4
n = 1000000, R = 10			05:45.2	05:24.5	05:06.0	04:49.5
n = 500000 , R = 20			05:26.4	05:17.0	05:06.7	05:00.8
n = 100000, R = 100			06:05.4	05:24.8	05:01.2	04:48.4
n = 50000, R = 200			05:30.8	05:15.5	05:43.4	05:24.5
n = 10000, R = 1000			06:16.4	04:47.0	07:07.5	05:42.1

Figure 4.3: Time taken by different procedures for 'Ten million' observations

choosing optimal 'R' and 'n' is crucial as otherwise it would fail to reduce time.

Now I revert back to the first important aspect of the results which come in the form of appropriateness of use of the estimators from the proposed procedures, viz. the simple mean estimator and the variance minimized estimator. In this paper, the parameter estimates found from P1 and P2 are compared to estimates from classical approach and the variances of these guide us to choose among P1 and P2. The results are again shown for the three different cases with observations one hundred thousand, one million and ten million. First, let us compare the results for our first dataset of 100000 observations. The model is again,

$$y_i = -2 + 0.8x_{i1} + 3.6x_{i2} + 1.3x_{i3} - 0.9x_{i4} + u_i \text{ where } i = 1(1)N \tag{4.1}$$

It is important to note here that the given R, n is so chosen such that nR = 100000. Now, from Figure 4.7 it is evident that there is not a significant difference between the parameter estimates and the standard errors from P2 are smaller than the standard errors obtained from P1, the simple means procedure. When R = 1, i.e., all the observations are selected with probability 1 in the replication, it gives the same parameter estimates and the standard errors. But, as R is increased to 10, the standard errors from P2 are not much different from the standard errors obtained from the classical approach while the standard errors from P1 are almost 10 times, i.e., the resulting variances would be almost 100 times the original variances for most of the parameters.

Similar results for the other two datasets are attached in the appendix. Since the time taken by both the procedures is almost the same and results show that for the two larger datasets, minimum time is achieved by P2 in both the cases and also variances are closer to the ones found by using the entire dataset. Hence P2 can be considered to produce reasonable estimators and better than the ones found from P1.

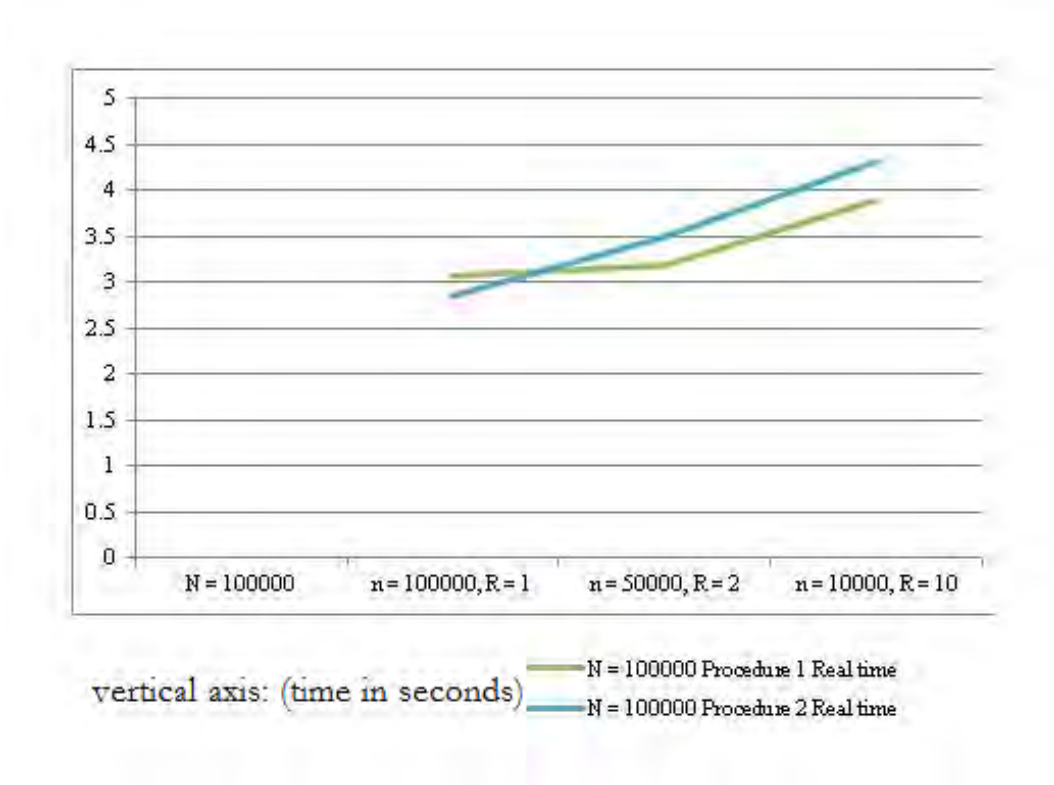

Figure 4.4: Plot of time against different replication numbers (R) and observations (n) in case of N =100000

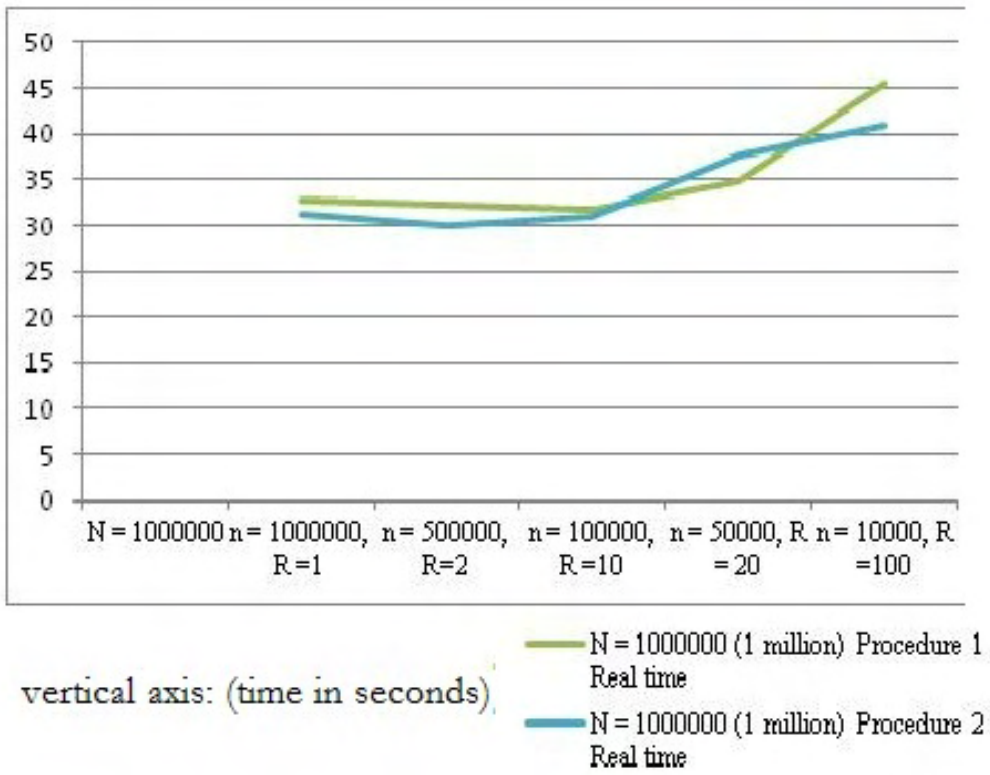

Figure 4.5: Plot of time against different replication numbers (R) and observations (n) in case of N = 1000000

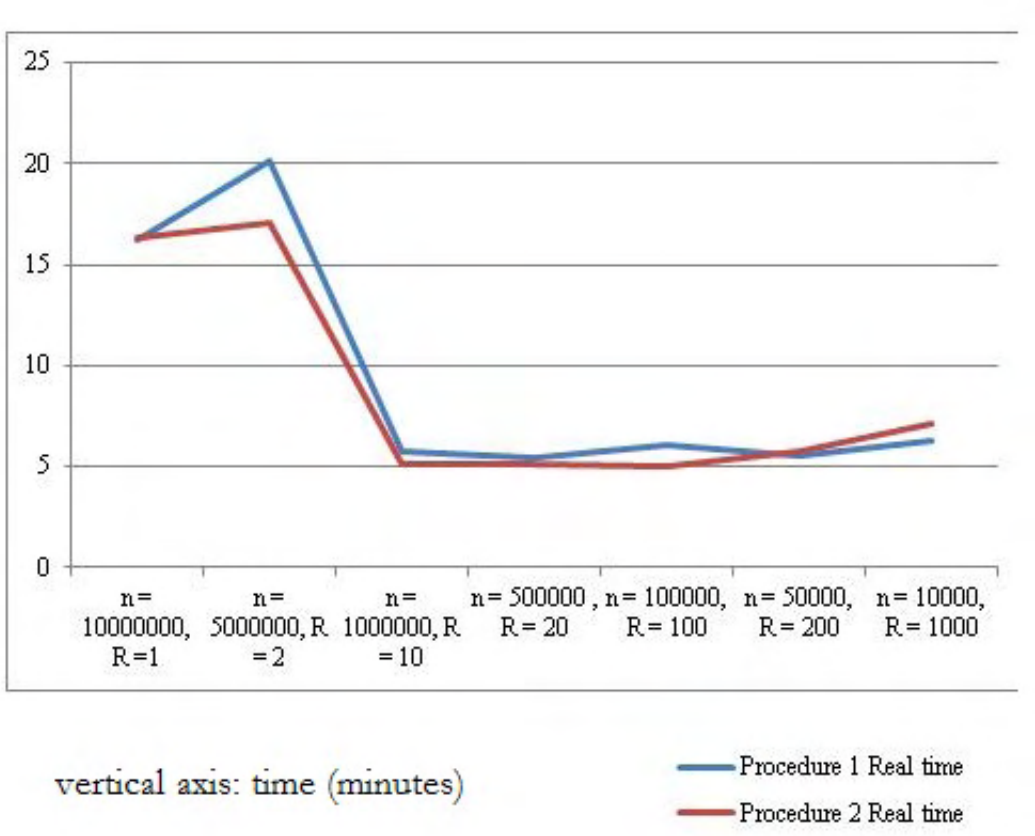

Figure 4.6: Plot of time against different replication numbers (R) and observations (n) in case of N = 10000000

		R = 1		R = 2		R = 10	
	All Observations	P1	P2	P1	P2	P1	P2
Intercept	-2.070588 (0.062820)	-2.0705881 (0.0628204)	-2.07059 (0.06282)	-2.0499593 (0.0889450)	-2.12633 (0.062860283)	-2.0766603 (0.1996138)	-2.06394 (0.062698)
β1	0.792334 (0.008359)	0.7923339 (0.0083595)	0.792334 (0.008359)	0.7908986(0.0118540)	0.790625 (0.008316165)	0.7956059 (0.0262796)	0.791936 (0.00838)
β2	3.633639 (0.039858)	3.6336394 (0.0398584)	3.633639(-0.039858)	3.6493353 (0.565345)	3.654688 (0.039876906)	3.6353792 (0.1259720)	3.631231 (0.039719)
β3	1.298779 (0.009587)	1.2987790 (0.0095870)	1.298779 (0.009587)	1.3111487 (0.0137454)	1.296051 (0.009564669)	1.2994789 (0.0303226)	1.299837 (0.009601)
β4	-0.883325 (0.019305)	-0.8833247 (0.0193053)	-0.88332 (0.019305)	-0.8840606 (0.0273761)	-0.88499 (0.01935087)	-0.8710988 (0.0609646)	-0.87528 (0.019381)

Figure 4.7: Comparing Estimates and Standard Errors from all procedures when N = 100000

Chapter 5

Conclusions

In the era of information technology advancement when massive datasets are no longer rare in the fields of research and science, business and finance, internet and retailing, statistical methods are available for data mining but not typically available for analysis of such huge datasets. While there is a substantial literature on how huge data can be analyzed for estimation in regression analysis of continuous variables, this paper is an attempt to come up with reasonable estimators in regression analysis of a probit model. The paper discusses about two proposed procedures in both of which the entire dataset is segregated in parts by randomly selecting observations by SRS technique and repeating the process to select replicates each having same number of observations. The first procedure P1 gives us a simple technique to aggregate the information from all replicates by simply taking the arithmetic mean of the parameter estimates from each replicate. It has been shown to be linear and unbiased. The variance for such an estimate is different from the arithmetic mean of variances from the replicates and hence is not a good proxy for the former. Finally, the second procedure P2 provides both theoretically and empirically results which give rise to linear unbiased estimators by minimizing the variance of the weighted estimate. This procedure is shown to be optimal in the sense of minimizing the variances of the estimates as well as reducing a significant amount of time in the estimation using randomly generated data.

While it has been shown using datasets of various sizes that as the data size grows larger, there is a significant amount of time that can be saved using both the procedures, but regression results show that variance minimized estimators have variances as low as the ones from the regression of full dataset. It has also been argued that the mean of variances are not a proxy for the variances of the simple means estimator and are hence P1 is a better procedure than P2 in terms of precision of proposed estimators. This proposed procedure gives estimators which can come very handy with extremely large datasets. Also when a researcher is choosing dependent variables to fit a probit model and hence has to run several regressions, the variance minimized estimators might be used instead of the classical estimator as it would significantly reduce the process time in presence of a very large data.

Further research is needed for applying these techniques in probit models with extremely large datasets. The asymptotic properties of these estimators are yet to be verified and it might be useful to see how these estimators behave if $nR < N$. Also, it might be a good exercise to compare these results to the ones in which the entire dataset is segregated into equal blocks by partitioning the data as has been done in the literature. The optimal number of blocks 'R' needs to be found out analytically for a dataset of a particular size given the specifications of a computer. Finally, since all the results pertaining to time depends a lot on the memory of a computer, the optimal size of R is supposed to vary for different specification and hence would require further research.

Appendix

Estimates	All Observations	R = 1		R = 2		R = 10	
		P1	P2	P1	P2	P1	P2
Intercept	-2.009807 (0.019891)	-2.0098075 (0.0198906)	-2.00981 (0.019891)	-2.0094396 (0.0281375)	-1.99073 (0.019869)	-2.0463413 (0.0629811)	-1.98447 (0.019892)
β1	0.799845 (0.002663)	0.07998451 (0.0026631)	0.799845 (0.002663)	0.799309 (0.0037679)	0.801295 (0.002668)	0.7973515 (0.0084187)	0.802064 (0.002665)
β2	3.615577 (0.012613)	3.6155769 (0.0126129)	3.615577 (0.012613)	3.6128246 (0.0178385)	3.60895 (0.012601)	3.6333021 (0.0399827)	3.597736 (0.012605)
β3	1.304934 (0.003045)	1.3049388 (0.0030446)	1.304934 (0.003045)	1.3041686 (0.0043013)	1.30712 (0.003054)	1.3054495 (0.0096410)	1.303028 (0.003039)
β4	-0.902021 (0.006141)	-0.9020213 (0.0061410)	-0.90202 (0.006141)	-0.8989643 (0.0086883)	-0.89899 (0.006149)	-0.8990535 (0.0194268)	-0.90223 (0.006129)
Estimates	All Observations	R = 20		R = 100			
		P1	P2	P1	P2		
Intercept	-2.009807 (0.019891)	-2.0326024 (0.0889559)	-2.00114 (0.019918)	-2.0260249 (0.1987843)	-2.01603 (0.019962)		
β1	0.799845 (0.002663)	0.08020620 (0.0119255)	0.80459 (0.002673)	0.8000450 (0.0266519)	0.799882 (0.00267)		
β2	3.615577 (0.012613)	3.6286047 (0.0564591)	3.612786 (0.012633)	3.6214307 (0.1261491)	3.630941 (0.012676)		
β3	1.304934 (0.003045)	1.3063709 (0.0136184)	1.306254 (0.003051)	1.3042758 (0.0304205)	1.311209 (0.003065)		
β4	-0.902021 (0.006141)	-0.8960651 (0.0274672)	-0.90532 (0.006152)	-0.9025270 (0.0614033)	-0.90566 (0.006166)		

Figure 5.1: Comparing estimates and standard errors from all procedures when N=1 million

Estimates	All Observations	R = 1		R = 2		R = 10	
		P1	P2	P1	P2	P1	P2
Intercept	-1.999906 (0.006276)	-1.9999055 (0.0062760)	-1.99991 (0.006276)	-1.9989444 (0.0088732)	-1.99535 (0.006272)	-2.0006619 (0.0198517)	-2.00743 (0.006282)
β1	0.799635 (0.000842)	0.7996346 (0.000841660)	0.799635 (0.000842)	0.7986973 (0.0011894)	0.79892 (0.000841)	0.7989769 (0.0026600)	0.798957 (0.000841)
β2	3.600356 (0.003972)	3.6003562 (0.0039723)	3.600356 (0.003972)	3.5996589 (0.0056157)	3.596599 (0.00397)	3.6008032 (0.0125595)	3.603646 (0.003974)
β3	1.300384 (0.000959)	1.3003842 (.000958681)	1.300384 (0.000959)	1.3003177 (0.0013553)	1.299421 (0.000958)	1.2999506 (0.0030298)	1.300037 (0.000958)
β4	-0.899286 (0.001937)	-0.8992859 (0.0019371)	-0.89929 (0.001937)	-0.8963961 (0.0027380)	-0.8983 (0.001937)	-0.9021757 (0.0061267)	-0.8988 (0.001936)
Estimates	All Observations	R = 20		R = 100		R = 200	
Intercept	-1.999906 (0.006276)	-1.9893310 (0.0280403)	-1.99959 (0.006278)	-2.0083775 (0.0627201)	-2.00316 (0.006283)	-1.9977638 (0.0887457)	-2.0011 (0.006277)
β1	0.799635 (0.000842)	0.7975807 (0.0037615)	0.800541 (0.00842)	0.7987144 (0.0084120)	0.800256 (0.000842)	0.8013604 (0.0119136)	0.80041 (0.000841)
β2	3.600356 (0.003972)	3.5950205 (0.0177411)	3.60173 (0.003973)	3.6011701 (0.0396956)	3.60089 (0.003976)	3.6013302 (0.0561821)	3.598852 (0.003972)
β3	1.300384 (0.000959)	1.3001563 (0.0042867)	1.301234 (0.000959)	1.2986568 (0.0095700)	1.299807 (0.000958)*	1.3018272 (0.0135788)	1.299845 (0.000958)
β4	-0.899286 (0.001937)	-0.9016815 (0.0086683)	-0.89827 (0.001937)	-0.8970771 (0.0193608)	-0.9009 (0.001937)	-0.8995967 (0.0274212)	-0.89644 (0.001936)
Estimates	All Observations	R = 1000					
Intercept	-1.999906 (0.006276)	-2.0055147 (0.1990646)	-2.00653 (0.006286)				
β1	0.799635 (0.000842)	0.08015248 (0.0266984)	0.799311 (0.000842)				
β2	3.600356 (0.003972)	3.6101350 (0.1260604)	3.604073 (0.003979)				
β3	1.300384 (0.000959)	1.3038322 (0.0304482)	1.300848 (0.00096)				
β4	-0.899286 (0.001937)	-0.9039255 (0.0614576)	-0.89534 (0.001939)				

Figure 5.2: Comparing estimates and standard errors from all procedures when N=10 million

* Note that the variance from P2 is less than that obtained from all observations due to rounding off after six decimals; otherwise it is statistically not possible. It also shows the proximity of the variances obtained by P2 when compared to the actual ones.

References

Balakrishnan, S. Madigan, D. A one-pass sequential Monte-Carlo method for Bayesian analysis of massive datasets, Bayesian Analysis 1 (2006) 345-362

Elder, J. Pregibon, D. A Statistical Perspective on Knowledge Discovery and Databases, AAAI/MIT Press, 1996 (Chapter 4)

Fan, T. Lin, D.K.J. Cheng, K.F. Regression analysis for massive datasets, Data Knowledge Engineering 61 (2007) 554-562

Glymour, C., Madigan, D., Pregibon, D., Smyth P., Statistical Themes and lessons for Data Mining, Data Mining and Knowledge Discovery, 1.1 (1997) 11-28

Hand, D.J. Blunt, G. Kelly, M.G. Adams, N.M. Data mining for fun and profit, Statistical Sciences 15 (2000) 111-131

Jackman S. Estimation and Inference via Bayesian Simulation: An introduction to Markov Chain Monte Carlo, American Journal of Political Science, 44. 2 (2000) 375-404

Li, R. Lin, D.K.J. Li, B. Statistical Inference on Large Data Sets, Knowledge Discovery

The Economist, 25 February 2010, Retrieved 9 December 2012, http://www.economist.com/node/15557443?story_id=15557443

Wikipedia: http://en.wikipedia.org/wiki/Big_data