AN ABSTRACT OF THE DISSERTATION OF

Olga Voyteshenko Livingston for the degree of Doctor of Philosophy in Economics presented on December 17, 2009.

Title: Essays in Semiparametric Econometrics.

Abstract approved:

Carlos Martins-Filho

Two essays are focused on semiparametric econometric methods. The first essay investigates applicability of the smooth backfitting estimator (SBE) to statistical analysis of residential energy consumption. The second essay attempts to incorporate additivity restrictions into semiparametric stochastic frontier estimation. The procedure described in the first study is used to estimate the directional regressions for each of the additive components. These estimates are used as a pilot for stochastic frontier estimation. The essay contains an empirical study of power-generating units in the US.

Essays in Semiparametric Econometrics

by

Olga Voyteshenko Livingston

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented December 17, 2009
Commencement June 2010

Doctor of Philosophy dissertation of Olga Voyteshenko Livingston presented on December 17, 2009.

APPROVED:

Major Professor, representing Economics

Chair of the Department of Economics

Director of the Economics Graduate Program

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Olga Voyteshenko Livingston, Author

TABLE OF CONTENTS

		Pages
1.	General introduction……	1
2.	Smooth backfitting estimation of natural gas consumption based on Residential Energy Consumption Survey microdata……	3
	2.1 Introduction……..	4
	2.2 Smooth backfitting for continuous data……	6
	2.3 Smooth backfitting estimator for mixed data……..	9
	2.4 Bandwidth selection……..	14
	2.5 Results and analysis……..	16
	2.6 Conclusion……..	27
3.	Semiparametric estimation of stochastic production frontier with additivity constraints……..	30
	3.1 Introduction……..	31
	3.2 Literature overview……	32

TABLE OF CONTENTS (Continued)

Pages

3.3 Model..... 34

3.3.1 Bandwidth selection..... 39
3.3.2 Estimation algorithm..... 42

3.4 Results and analysis..... 43

3.5 Conclusion..... 54

4. General conclusion..... 56

5. Bibliography..... 57

Appendix..... 62

Appendix A. Charts for directional regression results..... 63

LIST OF FIGURES

Figures		Pages
1.	Electricity output versus fixed O&M cost……	44
2.	Electricity output versus .variable O&M cost…….……....	44
3.	Electricity output versus cycle type……	45
4.	Electricity output versus cycle type……	45
5.	Smooth backfitting estimates for direction 1, fixed O&M cost…….….	46
6.	Smooth backfitting estimates for direction 2, variable O&M cost……..……..	46
7.	Smooth backfitting estimates for direction 3, fuel cost……..…….	47
8.	Smooth backfitting estimates for cycle type……..……	47
9.	Smooth backfitting estimates for combined cycle……..….	47
10.	Smooth backfitting estimates for combustion turbine……..……..	48
11.1	Smooth backfitting estimates for steam turbine, coal.....	48

LIST OF FIGURES (Continued)

Figures		Pages
11.2	Smooth backfitting estimates for steam turbine, gas……	48
11.3	Smooth backfitting estimates for steam turbine, other……	49
12.	Nonparametric density for unit-specific inefficiency……	49
13.	Nonparametric density estimates for technical efficiency…	50
14.	Scatter plot of TEI against the cycle types……..	51
15.	Scatter plot of unit- unit-specific efficiency ……	51
16.	95% Confidence intervals for unit-specific efficiency scores….....	54

LIST OF APPENDIX FIGURES

Figures		Pages
A.1	Heating degree days: base=65, 01 to 12-2005…….	63
A.2	Cooling degree days: base=65, 01 to 12-2005……..	63
A.3	Total house area……..……..	64
A.4	Price of electricity, cents/KWh……..……	64
A.5	Price of natural gas, cents*10/Btu……	65
A.6	Setting during the winter day when someone is home……..	65
A.7	Setting during the winter day when no one is home……..	66
A.8	Setting during sleeping hours in winter…….……	66
A.9	Exterior wall construction material…….....	67
A.10	Is the garage heated……	68
A.11	Dwelling owned or rented……	69
A.12	Fuel used by cooking burners…….	70

LIST OF APPENDIX FIGURES (Continued)

Figures		Pages
A.13	Fuel used by clothes dryer……..	71
A.14	Combined all secondary heating equipment……	72
A.15	Is the thermostat programmable……	73
A.16	Programmable thermostat lowers heat at night……	74
A.17	Programmable therm lowers heat during the day……..	75
A.18	Main fuel used for heating home……	76
A.19	Type of heating equipment providing the heat……	77
A.20	Natural gas used for water heating……..	78
A.21	How natural gas is paid……	79
A.22	Is someone at home all day on a typical weekday……..	80
A.23	Reported stories in housing unit……	81
A.24	Basement/crawl space heated……..	82

LIST OF APPENDIX FIGURES (Continued)

Figures		Pages
A.25	How much of the attic is warm……..	83
A.26	Year home built……..…	84
A.27	How many thermostats overall……..	85
A.28	Number of rooms not heated last winter……	86
A.29	Type of window glass……	87
A.30	Number of occupants (0=none, up to 10)……..	88
A.31	Total combined income in the past 12 months……	89

1 General introduction

This dissertation consists of two applied studies, which are focused on semiparametric methods. The first essay investigates the applicability of the smooth backfitting estimator (SBE) to statistical analysis of residential energy consumption via nonparametric regression. The methodology utilized in this study extends nonparametric additive regression via local linear smooth backfitting to categorical variables. This paper attempts to establish the relationship between energy demand and residential building attributes, demographic characteristics and behavioral variables using the Residential Energy Consumption Survey 2005 microdata. The computational algorithm developed in the first paper is then utilized in the second essay in a different setting. The second essay incorporates additivity restrictions into semiparametric stochastic frontier estimation. This study uses local linear smooth backfitting with categorical variables as a pilot estimator in the context of semiparametric stochastic frontier estimation of Fan et al. (1996), which is utilized to analyze efficiency of power generating units in the U.S. Both essays deal with current energy issues, and both studies employ nonparametric estimation procedures to analyze them.

The conventional methods used for analyzing residential energy consumption are econometric modeling and engineering simulations. The first paper suggests an econometric approach that can be utilized in combination with simulation results. A common weakness of previously used econometric models is a very high likelihood that any suggested parametric relationships will be misspecified. Nonparametric modeling does not have this drawback. Its flexibility allows for uncovering more complex relationships between energy use and the explanatory variables than can possibly be achieved by parametric models.

Traditionally, building simulation models overestimated the effects of energy efficiency measures when compared to actual "as-built" observed savings. While focusing on technical efficiency, they do not account for behavioral or market effects. The magnitude of behavioral or market effects may have a substantial influence on the final energy savings resulting from implementation of various energy conservation measures and programs. Moreover, variability in behavioral aspects and user characteristics appears to have a significant impact on total energy consumption. Inaccurate estimates of energy consumption and potential savings also impact investment decisions. The existing modeling literature, whether it relies on parametric specifications or engi-

neering simulation, does not accommodate inclusion of a behavioral component. The first paper attempts to bridge that gap and investigate the applicability of additive nonparametric regression to this task.

The second essay is also related to energy, but it is focused on the efficiency of power generation. The issues of energy security, reducing reliance on fossil fuels and reducing the carbon footprint have gained increased attention in recent literature. The feasibility of sufficiently meeting the growth in energy demand with renewable resources is still under discussion. Meanwhile, fossil fuels remain the dominant source of power generation in the United States. Current energy plans envision construction of additional coal plants, which is not consistent with the environmental goals that are gaining high visibility as a result of the ongoing research by various government and private institutions. This begs the question of whether coal plants are chosen as the most efficient technology among fossil-fueled power generation units. As such, the second essay analyzes efficiency of existing generation capacity with a focus on fossil fuels. We use categorical variables to account for different types of power generation cycles. In the context of frontier estimation, categorical variables are traditionally handled parametrically. This paper contributes to the existing frontier analysis literature by including kernel smoothing of the categorical variables as part of the estimation procedure and using smooth backfitting as a pilot estimator within the frontier estimation framework developed by Fan et al. (1996). If results are plausible, they can be used as a benchmark case to compare efficiency of the existing power-generating capacity under different policy scenarios aimed at curbing emissions and changing the manner in which the power generating industry is currently operating.

2 Smooth backfitting estimation of natural gas consumption based on Residential Energy Consumption Survey microdata

2.1 Introduction

There are three main approaches to residential energy demand analysis: engineering, socio-psychological and econometric. The engineering approach relies on simulating different types of building energy use within an engineering modeling framework such as Energy Plus, DOE-2 and the like, Crawley et al. (2004). These building energy simulation tools construct demand projections by performing hourly energy simulations of buildings, air-handling systems, and equipment based on building and weather characteristics and an assumed operation schedule. The second approach evaluates the impact of institutions, beliefs and group influences on the long-term trends in energy use. The econometric approach links energy use to prices of energy products and their substitutes, as well as household income, demographic characteristics and features of the occupied buildings. This essay fits into the third category exploring the behavioral aspects of energy consumption at the micro level.

Detailed studies of energy use at the household level using microeconomic data were conducted by Baker et al. (1989), Schmalensee and Stoker (1999), Halvorsen and Larsen (2001), Yatchew and No (2001), Nesbakken (2001) and Larsen and Nesbakken (2004), Garcia-Cerruti (2000), Holtedahl and Joutz (2004), Kamerschen and Porter(2004) and Narayan and Smyth (2005) to name a few. The reviewed econometric studies all estimate energy demand functions; however, the explanatory variables employed by these studies differ. These studies can generally be categorized into two groups. The first group includes economic variables such as fuel prices and income level, as well as climate information. The second group of studies incorporates additional household and demographic characteristics of the dwelling into the model. An extensive overview of econometric analysis of residential energy demand predating the above-listed research is included in Madlener (1996).

The focus of this study is residential natural gas (NG) demand. Space heating is the single largest end use of energy in residential buildings, and furnaces fueled by natural gas are the primary source of residential heating. Natural gas also provides fuel for residential water heating, cooking, clothes drying, and other miscellaneous uses. In terms of on-site energy use measured in British thermal units (Btu), in 2006 the Energy Information Administration (EIA) estimated that natural gas supplied

approximately 65% of 4.4 quadrillion Btu delivered for residential space heating, and approximately 68% of total residential site energy for water heating (DOE/EIA-0383,2009). The primary substitute for natural gas in residential homes is electricity (i.e., electric furnaces, heat-pumps, electric water heaters, etc.).

The majority of econometric research on electricity and natural gas consumption relies on a fully specified parametric functional relationship between energy use and its conditioning variables. As a result there is the potential for severe misspecification of the proposed econometric models. Also, the categorical variables, which are typically present in residential microdata, are usually treated either by including dummy variables or via sub-sample regression. For example, for treatment of educational level, geographical location of the home, ownership of the main dwelling and number of household members see Labandeira et al. (2004, 2006). Nonparametric modeling is robust to functional form misspecification. Its flexibility allows for uncovering more complex relationships between energy use and conditioning variables than can be possibly achieved by parametric models.

In this essay we adopt additive nonparametric modeling for energy consumption, which would be estimated using the smooth backfitting procedure of Mammen et al. (1999). This procedure achieves convergence rates equal to that of univariate models thus bypassing the curse of dimensionality. In addition, recognizing that both continuous and categorical variables impact energy demand, this application of backfitting procedure incorporates the kernel smoothing methods of Racine and Li (2003) and Racine et al. (2004) for categorical variables.

The smooth backfitting approach adopted in this essay is different from classical backfitting of Buja et al. (1989). The latter estimator is not efficient according to the "oracle efficiency" criteria put forth by Linton and Nielsen (1995), which means obtaining a directional regression estimator that is asymptotically the same as the estimator for which all other directions were known. It was shown by Opsomer and Ruppert (1997) and Opsomer (2000) that backfitting does not reach this oracle efficiency bound. Furthermore, the classical backfitting estimator is not known to be asymptotically normal.

Smooth backfitting is oracle efficient. Moreover, it has the intuitive geometrical interpretation of a projection of the data onto the space of additive functions. Smooth backfitting possesses a high degree of implementational appeal as its iterative equations rely on the estimates of univariate regressions for each covariate, as well as

univariate and bivariate densities only. The theoretical properties of the estimator were also derived in Mammen et al. (1999) who showed the conditions for convergence and uniqueness of the estimator. In addition, smooth backfitting is capable of satisfactorily accommodating covariates with significant degrees of correlation as demonstrated by Nielsen and Sperlich (2005).

The data for this research comes from the Residential Energy Consumption Survey designed by the US Department of Energy's Energy Information Administration. The microdata obtained from the 2005 survey covers energy consumption for several major fuel types and includes information on household characteristics, standard demographics, dwelling characteristics, as well as information about televisions and other media devices, personal computers and peripherals, Energy Star labeling, energy efficient lighting, window glazing, window replacement, and thermostat usage. The 2005 survey also incorporates questions on behavioral aspects of energy use. This analysis contributes to existing literature by analyzing and quantifying behavioral impacts on residential energy consumption.

The essay is organized into five sections. A brief description of the local linear smooth backfitting estimator (SBE) for continuous variables is presented in Section 2.2. Section 2.3 contains the extension of the local linear smooth backfitting estimator to mixed variables. Section 2.4 includes an explanation of the bandwidth selection procedure. Section 2.5 describes the results of the empirical analysis. Charts for regression results in each direction, as well as the list of variables with category labels for unordered and ordered discrete variables are included in Appendix A.

2.2 Smooth backfitting for continuous data

The regression model considered here is of the following form:

$$E(Y|X_1 = x_1, ..., X_d = x_d) = m_0 + \sum_{j=1}^{d} m_j(x_j)$$

where $(Y, X_1, ..., X_d)$ is a random vector in $\mathbb{R}^{d+1}$ and we assume that there is a random sample $\{y_i, x_{i1}..., x_{id}\}_{i=1}^{n}$ of $(Y, X_1..., X_d)$, m_0 is an unknown scalar parameter, $m_j(x_j)$ is a sufficiently smooth function for all j, and θ_j is the first order derivative of $m_j(x_j)$. Also, for identification purposes, $E(m_j(x_j)) = 0$.

Let $K_h(x_{ij}-x_j)=\frac{1}{h}K\left(\frac{x_{ij}-x_j}{h}\right)$ be a kernel function such that $\int K(\phi)\,d\phi=1$, $\int \phi K(\phi)\,d\phi=0$, $\int \phi^2 K(\phi)\,d\phi=1$. Bandwidth is defined as $h=h(n)$ such that $h\to 0$ and $nh\to\infty$ as $n\to\infty$, and conditions B(1), B(2')-B(4') of Mammen et al. (1999) are met. The backfitting estimator is obtained by minimizing the following objective function

$$\int \sum_{i=1}^{n}\left[y_i-m_0-\sum_{j=1}^{d}m_j(x_j)-\sum_{j=1}^{d}\theta_j(x_j)(x_{ij}-x_j)\right]^2\times\prod_{j=1}^{d}K_h(x_{ij}-x_j)\,dx$$

The minimization is done with respect to m_0, $m_1...m_d$ and all first derivatives $\theta_j(x_j)$.

Let

$$\widehat{p}_j(x_j)=n^{-1}\sum_{i=1}^{n}K_h(x_{ij}-x_j),\quad \widehat{p}_j^{\,j}(x_j)=n^{-1}\sum_{i=1}^{n}K_h(x_{ij}-x_j)(x_{ij}-x_j),$$

$$\widehat{p}_j^{\,jj}(x_j)=n^{-1}\sum_{i=1}^{n}K_h(x_{ij}-x_j)(x_{ij}-x_j)(x_{ij}-x_j),$$

$$\widehat{p}_{jk}(x_j,x_k)=n^{-1}\sum_{i=1}^{n}K_h(x_{ij}-x_j)K_h(x_{ik}-x_k),$$

$$\widehat{p}_{jk}^{\,k}(x_j,x_k)=n^{-1}\sum_{i=1}^{n}K_h(x_{ij}-x_j)K_h(x_{ik}-x_k)(x_{ik}-x_k),$$

$$\widehat{p}_{jk}^{\,jk}(x_j,x_k)=n^{-1}\sum_{i=1}^{n}K_h(x_{ij}-x_j)K_h(x_{ik}-x_k)(x_{ij}-x_j)(x_{ik}-x_k),$$

Let

$$\begin{aligned}A &= \frac{n^{-1}\sum_{i=1}^{n}K_h(x_{ij}-x_j)\,y_i}{\widehat{p}_j(x_j)}-\sum_{k\neq j}^{d}\int\widetilde{m}_k(x_k)\frac{\widehat{p}_{jk}(x_j,x_k)}{\widehat{p}_j(x_j)}dx_k\\ &\quad-\sum_{k\neq j}^{d}\int\widetilde{\theta}_k(x_k)\frac{\widehat{p}_{jk}^{\,k}(x_j,x_k)}{\widehat{p}_j(x_j)}dx_k-\widetilde{m}_0(x_j),\\ B &= \frac{n^{-1}\sum_{i=1}^{n}K_h(x_{ij}-x_j)(x_j-X_{ij})\,y_i}{\widehat{p}_j^{\,j}(x_j)}-\sum_{k\neq j}^{d}\int\widetilde{m}_k(x_k)\frac{\widehat{p}_{jk}^{\,j}(x_j,x_k)}{\widehat{p}_j^{\,j}(x_j)}dx_k\\ &\quad-\sum_{k\neq j}^{d}\int\widetilde{\theta}_k(x_k)\frac{\widehat{p}_{jk}^{\,jk}(x_j,x_k)}{\widehat{p}_j^{\,j}(x_j)}dx_k-\widetilde{m}_0(x)\end{aligned}$$

$$C = \frac{\widehat{p}_j^{j}(x_j)}{\widehat{p}_j(x_j)}, \qquad D = \frac{\widehat{p}_j^{jj}(x_j)}{\widehat{p}_j^{j}(x_j)}$$

The smooth backfitting estimates of $\widetilde{m_0}$, $\widetilde{m_j}$ and $\widetilde{\theta}_j$ are obtained by iteratively solving the two equations below for each regressor $j = 1, ..., d$

$$\widetilde{m_j}(x_j) = A - \widetilde{\theta}_j(x_j)C, \qquad \widetilde{\theta}_j(x_j) = \frac{A - B}{C - D}$$

As a consequence of imposing normalization condition, $\widetilde{m_0} = n^{-1}\sum_{i=1}^{n} y_i$.

A detailed discussion establishing the asymptotic properties of the smooth backfitting estimator for the case of only continuous regressors is presented in Mammen et al. (1999). Their final result is summarized as the convergence in distribution that holds for any $x_1, ...x_d$ with compact support:

$$n^{2/5}\begin{pmatrix} \widetilde{m}_1(x_1) - m_1(x_1) + v_{n,1} \\ \vdots \\ \widetilde{m}_d(x_d) - m_d(x_d) + v_{n,d} \end{pmatrix} \xrightarrow{d} N\left[\begin{pmatrix} c_h^2\delta_1(x_1) \\ \vdots \\ c_h^2\delta_d(x_d) \end{pmatrix}, \operatorname{diag}\left\{\upsilon_j(x_j)\right\}_{j=1}^{d}\right],$$

$$\delta_j(x_j) = \frac{\int u^2 K(u)du}{2}\left\{m_j{}''(x_j) - \int m_j''(x_j)p_j(x_j)dx_j\right\},$$

$$v_{n,j} = \int m_j(x_j)K_h(x_j - u)p_j(u)du\ dx_j,$$

$$\upsilon_j(x_j) = c_h^{-1}c_k\sigma_j^2(x_j)/p_j(x_j),$$

with $c_k = \int K(u)^2du$, c_h is a constant such that $n^{1/5}h \to c_h$. Second derivative of $m_j(x_j)$ is represented by $m_j{}''(x_j)$, $p_j(u)$ is the marginal density, and $\sigma_j^2(x_j) = var[Y - m(x)|X_j = x_j]$ can be consistently estimated from the residuals $\widetilde{\varepsilon}_i = y_i - \widetilde{m}(x_i)$, $i = 1...n$.

$$n^{2/5}\left(\widetilde{m}(x) - m(x)\right) \xrightarrow{d} N\left\{c_h^2\sum_{j=1}^{d}\delta_j(x_j),\ \sum_{j=1}^{d}\upsilon_j(x_j)\right\},$$

where $\widetilde{m}(x)$ is a smooth backfitting estimator of $m(x) = m_0 + \sum_{j=1}^{d} m_j(x_j)$ defined as $\widetilde{m}(x) = \widetilde{m_0} + \sum_{j=1}^{d}\widetilde{m_j}(x_j)$.

2.3 Smooth backfitting estimator for mixed data

In a wide variety of applications, especially dealing with microdata, one of the essential features of a regression estimator is its capability to accommodate continuous and categorical conditioning variables. Traditional approaches for estimating the categorical components have relied either on introducing these variables parametrically or implementing a frequency-based estimation. The major drawback of the first approach is a loss of flexibility induced by a fully nonparametric framework, as well as high likelihood of misspecification. The weakness of the second method stems from the requirement to divide the data into cells corresponding to the values taken by the discrete variables. This necessitates fairly large sample size in order for each cell to contain a reasonable amount of data as described in Li and Racine (2007).

Alternative procedures, such as smooth estimation of joint distributions and smooth regression for discrete data, are based on kernel estimation proposed by Aitchison and Aitken (1976). This latter method received attention in the recent literature as kernel smoothing methods have been gaining popularity. Li and Racine (2003) proposed a refined nonparametric kernel approach for estimating an unknown distribution defined over mixed discrete and continuous variables. Nonparametric estimation of regression functions was investigated by Racine and Li (2004), where specific smoothing techniques were considered for treatment of ordered and unordered categorical data. Structure of the proposed estimator is similar to that of Nadaraya-Watson local constant estimator, but with a different kernel employed for smoothing discrete variables. Li and Racine (2004) expanded the regression framework further by constructing a local linear nonparametric estimator for mixed data and investigating the theoretical properties of cross-validated bandwidth selection. In addition, they derived the rate of convergence of the cross-validated bandwidths and established asymptotic normality of the resulting nonparametric regression estimator. These results provide a foundation for incorporating categorical regressors into the local linear smooth backfitting estimator (SBE) and using least squares cross-validation to select bandwidth for both continuous and categorical regressors.

Let x_j, $j = 1, ..., d$, denote continuous regressors and x_t, $t = 1, ...T$ denote the categorical variables. Discrete x_{it}, $i = 1, ...n$, takes values $\{0, 1, 2, ..., c_t - 1\}$. For the local linear regression estimator Li and Racine (2004) propose using a variation

of the Aitchison and Aitken (1976) kernel defined as

$$L(x_{it}, x_t, \lambda_t) = \begin{cases} 1, & \text{if } x_{it} = x_t \\ \lambda_t, & \text{if } x_{it} \neq x_t \end{cases} \quad t = 1, ...T.$$

This weight function does not add up to one, which cannot support the interpretation of marginal density $p_t(x_t)$ estimated by $\widehat{p}_t(x_t) = n^{-1} \sum_{i=1}^{n} L\ (x_{it}, x_t, \lambda_t)$ as a proper density. It has been shown by Li and Racine (2004) that it is not the kernel shape, but rather the selection of the bandwidth parameter that has critical impact on the quality of resulting estimates. Therefore, to accommodate interpretation of weighting functions in smooth backfitting estimation as densities, another option is to use the kernel shape suggested by Aitchison and Aitken (1976) for the distribution estimation, namely

$$L(x_{it}, x_t, \lambda_t) = \begin{cases} 1 - \lambda_t, & \text{if } x_{it} = x_t \\ \lambda_t / (c_t - 1), & \text{if } x_{it} \neq x_t \end{cases} \quad t = 1, ...T$$

for unordered categorical regressors. The range of λ_t is $[0, (c_t - 1)/c_t]$. This weight function adds up to one. When λ_t assumes its upper value of $(c_t - 1)/c_t$, the kernel becomes $L(x_{it}, x_t, \lambda_t) = 1/c_t$ regardless of whether $X_{it} = x_t$ or not. The resulting density estimator becomes unrelated to x_t thus smoothing it out. Alternatively, it is possible to use the weighting function that does not add up to one along with the normalization $p = p_t(x_t) / \sum p_t(x_t)$. For ordered categorical variable x_t the kernel of Li and Racine (2004)

$$L(x_{it}, x_t, \lambda_t) = \begin{cases} 1, & \text{if } x_{it} = x_t \\ \lambda_t^{|x_{it} - x_t|}, & \text{if } x_{it} \neq x_t \end{cases}$$

is utilized along with the above-mentioned normalization. The range of λ_t for ordered variables is [0,1]. If λ_t takes its upper value the kernel becomes a uniform weight function. If $\lambda_t = 0$, the kernel turns into an indicator function. An alternative is to use the kernel

$$L(x_{it}, x_t, \lambda_t) = \begin{cases} 1 - \lambda_t, & \text{if } |x_{it} - x_t| = 0 \\ \frac{1 - \lambda_t}{2} \lambda_t^{|x_{it} - x_t|}, & \text{if } |x_{it} - x_t| \geq 1 \end{cases},$$

where x_t is a categorical variable and x_{it}, $i = 1, ...n$, takes values $\{0, 1, 2, ..., c_t - 1\}$,

as proposed by Wang and van Ryzin (1981).

The multivariate discrete data kernel is defined as $\prod_{t=1}^{T} L(x_{it}, x_t, \lambda_t)$, with joint density of discrete variables being estimated by $\widehat{p}(x_1, ...x_T) = n^{-1} \sum_{i=1}^{n} \prod_{t=1}^{T} L(x_{it}, x_t, \lambda_t)$. The multivariate kernel for mixed data is

$$W\left(x_{ij}, x_j, h, x_{it}, x_t, \lambda_t\right) = \sum_{i=1}^{n} \prod_{j=1}^{d} K_h\left(x_{ij} - x_j\right) \prod_{t=1}^{T} L(x_{it}, x_t, \lambda_t).$$

The local linear estimator for continuous and discrete data suggested by Li and Racine (2004) has the following structure:

$$\begin{aligned} \widehat{s}(x) &= \begin{bmatrix} \widehat{m}(x) \\ \widehat{\theta}(x) \end{bmatrix} = \left[\sum_{i=1}^{n} W(x_{ij}, x_j, h, x_{it}, x_t, \lambda_t) \begin{pmatrix} 1 & (x_{ij} - x_j) \\ (x_{ij} - x_j) & (x_{ij} - x_j)^2 \end{pmatrix} \right]^{-1} \\ &\times \sum_{i=1}^{n} W\left(x_{ij}, x_j, h, x_{it}, x_t, \lambda_t\right) \begin{pmatrix} 1 \\ (x_{ij} - x_j) \end{pmatrix} y_i, \end{aligned}$$

where $s(x) = (m(x), \theta(x)')'$, $\theta(x) = \nabla\theta(x) = [\partial m(x)/\partial x_1, ...\partial m(x)/\partial x_d]'$. The partial derivative is taken only with respect to continuous variables. This estimator has the local constant shape for the discrete variables and local linear shape for the continuous variables.

The local linear smooth backfitting estimator for mixed continuous and categorical data is a projection of the local linear estimator for mixed regressors onto the space of additive functions. The mixed data local linear smooth backfitting estimator $\widetilde{m}^*(x)$ is defined as the argument that minimizes the following objective function

$$\begin{aligned} &\int \sum_{i=1}^{n} \left[y_i - m_0 - \sum_{j=1}^{d} m_j(x_j) - \sum_{t=1}^{T} m_t(x_t) - \sum_{j=1}^{d} \theta_j\left(x_{ij} - x_j\right) \right]^2 \\ &\times \prod_{j=1}^{d} K_h\left(x_{ij} - x_j\right) \prod_{t=1}^{T} L(x_{it}, x_t, \lambda_t) dx, \end{aligned}$$

where the categorical regressors are indexed by t. Derivation of the first order conditions for this setting follows the same logic as for the continuous regressors, where the minimization is performed over $m_0, m_j(x_j)$and $m_t(x_t)$ while preserving mean zero restriction, and over $\theta_j(x_j)$ for the continuous components only.

Using similar notation as before

$$\begin{aligned}\widetilde{m_j}(x_j) &= \frac{n^{-1}\sum_{i=1}^{n} K_h\left(x_{ij}-x_j\right) y_i}{\widehat{p}_j(x_j)} - \sum_{k\neq j}^{d}\int \widetilde{m_k}(x_k)\frac{\widehat{p}_{jk}(x_j,x_k)}{\widehat{p}_j(x_j)}dx_k \\ &\quad -\sum_{t=1}^{T}\int \widetilde{m_t}(x_t)\frac{\widehat{p}_{jt}(x_j,x_t)}{\widehat{p}_j(x_j)}dx_t - \sum_{k\neq j}^{d}\int \widetilde{\theta}_k(x_k)\frac{\widehat{p}^{k}_{jk}(x_j,x_k)}{\widehat{p}_j(x_j)}dx_k \\ &\quad -\widetilde{m_0}(x) - \widetilde{\theta}_j(x_j)\frac{\widehat{p}^{j}_{j}(x_j)}{\widehat{p}_j(x_j)},\end{aligned}$$

$$\begin{aligned}\widetilde{m_j}(x_j) &= \frac{n^{-1}\sum_{i=1}^{n} K_h\left(x_{ij}-x_j\right)\left(x_{ij}-x_j\right) y_i}{\widehat{p}^{j}_j(x_j)} - \sum_{k\neq j}^{d}\int \widetilde{m_k}(x_k)\frac{\widehat{p}^{j}_{jk}(x_j,x_k)}{\widehat{p}^{j}_j(x_j)}dx_k \\ &\quad -\sum_{t=1}^{T}\int \widetilde{m_t}(x_t)\frac{\widehat{p}^{j}_{jt}(x_j,x_t)}{\widehat{p}^{j}_j(x_j)}dx_t - \sum_{k\neq j}^{d}\int \widetilde{\theta}_k(x_k)\frac{\widehat{p}^{jk}_{jk}(x_j,x_k)}{\widehat{p}^{j}_j(x_j)}dx_k \\ &\quad -\widetilde{m_0}(x) - \widetilde{\theta}_j(x_j)\frac{\widehat{p}^{jj}_{j}(x_j)}{\widehat{p}^{j}_j(x_j)},\end{aligned}$$

where $\widetilde{m_0}(x)$ is the same as in continuous SBE setting. The iterative equations are shown below:

$$\begin{aligned}\widetilde{m^*_j}(x_j) &= A - \sum_{t\neq j}^{T}\int \widetilde{m_t}(x_t)\frac{\widehat{p}_{jt}(x_j,x_t)}{\widehat{p}_j(x_j)}dx_t - \widetilde{\theta^*_j}(x_j)C \\ &= A^* - \widetilde{\theta^*_j}(x_j)C\end{aligned}$$

$$\begin{aligned}\widetilde{m^*_j}(x_j) &= B - \sum_{t\neq j}^{T}\int \widetilde{m_t}(x_t)\frac{\widehat{p}^{j}_{jt}(x_j,x_t)}{\widehat{p}^{j}_j(x_j)}dx_t - \widetilde{\theta^*_j}(x_j)D \\ &= B^* - \widetilde{\theta^*_j}(x_j)D\end{aligned}$$

$$\widetilde{\theta^*_j}(x_j) = \frac{A^*-B^*}{C-D}.$$

Iterative equation for discrete regressors x_t, $t = 1, ..., T$ is

$$\begin{aligned}\widetilde{m_t^*}(x_t) &= \frac{\sum_{i=1}^{n} L\ (x_{it}, x_t, \lambda_t) y_i}{\widehat{p}_t(x_t)} - \sum_{j=1}^{d} \int \widetilde{m_j}(x_j) \frac{\widehat{p}_{jt}(x_j, x_t)}{\widehat{p}_t(x_t)} dx_j - \widetilde{m_0}(x) \\ &\quad - \sum_{k \neq t}^{T} \int \widetilde{m_k}(x_k) \frac{\widehat{p}_{kt}(x_k, x_t)}{\widehat{p}_t(x_t)} dx_k - \sum_{j=1}^{d} \int \widetilde{\theta}_j(x_j) \frac{\widehat{p}^j_{jt}(x_j, x_t)}{\widehat{p}_t(x_t)} dx_j.\end{aligned}$$

The last four equations jointly with the zero-mean condition describe the solution. Analogously to the continuous regressor densities

$$\widehat{p}_t(x_t) = n^{-1} \sum_{i=1}^{n} L\ (x_{it}, x_t, \lambda_t),$$

$$\widehat{p}_{jt}(x_j, x_t) = n^{-1} \sum_{i=1}^{n} K_h\left(x_{ij} - x_j\right)\ L\ (x_{it}, x_t, \lambda_t),$$

$$\widehat{p}^j_{jt}(x_j, x_t) = n^{-1} \sum_{i=1}^{n} K_h\left(x_{ij} - x_j\right)\ L\ (x_{it}, x_t, \lambda_t)\left(x_{ij} - x_j\right).$$

The algorithm for computation is as follows:

1. Compute the univariate $\widehat{p}_j(x_j)$, $\widehat{p}_t(x_t)$ for all regressors x_j and x_t, $j = 1, ...d$, and $t = 1, ...T$; compute $\widehat{p}^j_j(x_j)$, $\widehat{p}^{jj}_j(x_j)$ only for continuous components. Compute bivariate densities.

2. Compute univariate unrestricted $\widehat{m_t}(x_t) = (\sum_{i=1}^{n} L\ (x_{it}, x_t, \lambda_t) y_i) \diagup \widehat{p}_t(x_t)$ for all discrete variables and pairs $(\widehat{m}_j(x_j), \widehat{\theta}_j(x_j))$ for all continuous data. Save the results as variables m_{old} and θ_{old}.

3. Set the number of smooth backfitting iteration *iter* to 1.

   (a) For $j = 1$ compute expressions A*, B*, C, D. Obtain $\widetilde{m^*_j}(x_j)$ and $\widetilde{\theta^*_j}(x_j)$, save as m_{new} and θ_{new}. Repeat this step for the rest of continuous variables $j = 2, ...d$. To compute expressions A* and B*, use updated values from m_{new} and θ_{new} for $k < j$. If $k > j$, use corresponding values from m_{old} and θ_{old}.

   (b) Perform computation for discrete variables in a similar manner, with the conditional mean of categorical x_k in A being taken only over unique categories of x_k.

4. Define a convergence criteria for all j as $\frac{\sum_{i=1}^{n}\left[\widetilde{m_j^{new}}(x_j)-\widetilde{m_j^{old}}(x_j)\right]^2}{\sum_{i=1}^{n}\left[\widetilde{m_j^{old}}(x_j)\right]^2+\epsilon} < \epsilon$.

5. Set $iter = iter + 1$, Set m_{old} =m_{new} and θ_{old}=θ_{new}, then go to step 3a. Iterate steps 3a through 5 until the convergence criteria is met.

If $\int \widehat{p_{j,k}^{j,k}}(x_j, x_k)dx_k = \widehat{p_{j,k}^{j}}(x_j)$ does not hold, it is necessary to include the norming for $\widetilde{m_j^*}(x_j)$ such that $\widetilde{m_j^{*,n}}(x_j) = \widetilde{m_j^*}(x_j) - \int \widetilde{m_j^*}(x_j)\widehat{p_j}(x_j)dx_j$ after every iterative step for each $j = 1, ...T$. When the value of overall sum $m_0 + \sum_{j=1}^{d} m_j(x_j) + \sum_{t=1}^{T} m_t(x_t)$ is the primary point of interest, this normalization could be omitted as suggested in Mammen et al. (1999).

2.4 Bandwidth selection

Several different methods for selecting bandwidths for SBE estimation were analyzed recently. Mammen and Park (2005) introduced a bandwidth selection method for smooth backfitting based on minimizing penalized sum of squares residuals. They also compared two additional plug-in methods for local linear SBE. It was suggested that the penalized sum of squared residuals was asymptotically equivalent to cross-validation since this holds true for the classical nonparametric regression as in Hardle et al. (1988).

Leave-one-out least squares cross-validation is recommended for bandwidth selection by Nielsen and Sperlich (2005). It has an implementation advantage for local linear smooth backfitting if the underlying relationship is additive. In this case the cross-validation procedure can be simplified since the SB estimator has additively separable bias and variance. Bandwidth selection is based on minimizing mean-integrated squared error $MSE(h_1, ...h_d, \lambda_1, ...\lambda_d) = \int E\left[\widetilde{m}(x) - m(x)\right]^2 p(x)dx$. Due to separability of bias and variance, the mean-integrated squared error for overall regression can be defined as

$$MSE(h_1, ...h_d, \lambda_1, ...\lambda_d) = \sum_{j=1}^{d+T} MSE_j(x_j),$$

where $MSE_j(x_j)$ is mean-integrated squared error for each regression direction $m_j(x_j)$. Thus, the cross-validation problem of minimizing $CV = \sum_{i=1}^{n}\left[y_i - \widetilde{m}^{-i}(x)\right]^2$, where

$\widetilde{m}^{-i}(x)$ is the leave-one-out estimator with observation (y_i, x_i) excluded from the computation, can be separated. It reduces to performing an optimal bandwidth search for each directional regression sequentially. Nielsen and Sperlich (2005) suggest taking starting bandwidths $h_1, ...h_d$ that undersmooth for each direction and running the initial SBE estimation. Then the cross-validation criteria is minimized with respect to h_j only, where h_j is the bandwidth for direction j, by using a one-dimensional grid search. Bandwidths for all other directions are kept at their starting values. This is repeated for each direction j individually. It is not necessary to use leave-one-out estimators for all other directions $m_k(x_k)$, $k \neq j$, while searching for the optimal bandwidth for the estimation of $m_j(x_j)$. In addition, all $\widetilde{m_k}(x_k)$ do not need to be estimated at their optimal bandwidth. As shown by Mammen and Park (2005), this procedure results in bandwidths that are optimal for the estimation of the overall regression. If the primary focus of the estimation is accuracy of each single additive component, Mammen and Park (2005) suggest using plug-in bandwidths that minimize average weighted squared error (ASE) for each direction defined as

$$ASE_j(x_j) = n^{-1} \sum_{i=1}^{n} w_j^{-i}(x_j) \left[\widetilde{m_j}(x_j) - \widetilde{m}_j^{-i}(x_j)\right]^2,$$

where $\widetilde{m}_j^{-i}(x_j)$ is the leave-one-out estimator of $m_j(x_j)$ and w_j is a weight function.

This essay adopts a simpler method for bandwidth selection. Since smooth backfitting requires computing the unrestricted regression estimates, as well as univariate and bivariate densities for continuous and categorical data, we use four different bandwidth selection routines. To estimate densities for categorical variables we use the cross-validation method of Li and Racine (2007), where the bandwidth λ is chosen separately for each regressor to minimize

$$CV_p(\lambda) = \sum_{x_c \in S_c} \left[\widehat{p}\,(x_c)\right]^2 - 2n^{-2} \sum_{i=1}^{n} \sum_{v \neq i}^{n} L_{\lambda,iv},$$

where $L_{\lambda,iv}$ is the previously defined kernel with observation $v = i$ excluded from the computation, $S_c = \{0, ...c_t - 1\}$ is the support of x_c and c is the category index. For unrestricted regression estimation for categorical variables, the cross-validation of Li

and Racine (2007) is employed. Bandwidth is chosen to minimize

$$CV_{reg}(\lambda) = n^{-1} \sum_{i=1}^{n} \left[y_i - \widehat{m_j}^{-i}(x_j) \right]^2$$

for each j, where $\widehat{m_j}^{-i}(x_j)$ is the leave-one-out Nadaraya-Watson estimator of $m_j(x_j)$ defined as $\widehat{m_j}^{-i}(x_j) = \sum_{v \neq i}^{n} y_v \ L_{\lambda,iv} \Big/ \sum_{v \neq i}^{n} L_{\lambda,iv}$ For continuous variables the rule-of-thumb bandwidth selection was used both for estimation of unrestricted univariate regression, as well as densities. Namely, the bandwidth for regression estimation was selected as

$$h_j^{reg} = n^{-1/5} \left\{ s^2 2\sqrt{\pi} \left(\max(x_j) - \min(x_j) \right) \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \left(\widehat{b_3} + \widehat{b_4} x_j + 0.5 \widehat{b_5} x_j^2 \right)^2 \right]^{-1} \right\}^{1/5},$$

where b_3, b_4 and b_5 are estimates of coefficients in regressing the dependent variable y on $\beta_1 + \beta_2 x_j + \beta_3 (0.5 x_j^2) + \beta_4 (\frac{1}{6} x_j^3) + \beta_5 (\frac{1}{24} x_j^4)$, and s^2 is estimated in a usual manner based on the residual estimates of this regression. The bandwidth for density estimation was computed as $hdens_j = (n^{-1/5}) \cdot 1.01 a \left(2\sqrt{\pi}\right)^{-1/5}$, and $a = q_{75}(x_j) - q_{25}(x_j)$, where q_{75} and q_{25} are upper and lower quartiles of x_j, correspondingly.

2.5 Results and analysis

Upon close examination of the Residential Energy Consumption survey (RECS) questions and microdata for 2005, it became apparent that it would be an extremely complex task to cover all the end fuel uses for all fuel types included in the survey. The decision was made to investigate the applicability of smooth backfitting by isolating natural gas usage and related variables. The regressand is natural gas usage in British thermal units (Btu). There are 31 regressors that enter the model additively, 8 of which are continuous variables, 14 are unordered categorical variables and the remaining 9 are ordered categorical variables. A complete list of variables is included in Appendix A. Categories for each variable with corresponding regressor values are listed there as well. Individual crossvalidated bandwidth values were computed for each regressor. Initially the model was to include 44 categorical variables, but cross-validation produced the bandwidth values equal to the upper bound of $(c_t - 1)/c_t$ for

13 of the categorical variables. As mentioned in the bandwidth selection discussion, when the bandwidth takes this upper value it implies that the regressor is irrelevant and, if included, it will effectively be smoothed out. So these 13 variables were excluded from the analysis.

The charts with regression values for each direction are shown in Appendix A. Each regression direction is labeled accordingly. The first 8 directions represent continuous variables. The remaining variables are categorical. For categorical variables there are no values corresponding to the intervals between each threshold. The lines are included only for simplicity of illustration. Individual bandwidth was used for each of the directions.

Direction 1, heating degree days, seems to correctly represent the increase in natural gas intensity as the number of heating degree days goes up. Heating degree days are a characterization of weather. It is worth noting that RECS microdataset has sanitized data for heating and cooling degree days to prevent identification of survey respondents or specific buildings out of the reported sample. Even with the sanitized data the overall pattern of dependency is reasonable. Annual heating degree-days (HDD) are a measure of how cold a building location is relative to the base temperature. The daily HDD is the numerical difference between a day's average temperature and 65 degrees, if the average temperature is less than 65. Otherwise it is zero. Annual HDD is the sum of the daily HDD for the year. If the thermal integrity (e.g. insulation levels) of the building is known, it is possible to assess heating requirements from this information. The suggested pattern follows the engineering results that building heating requirements are not linear with respect to temperature. Therefore, natural gas use for heating will also have non-linear dependency on temperature. Although this pattern of dependency is well-known from engineering studies, the primary reason for including this variable is to analyze impact of other factors on energy demand, while controlling for weather.

Direction 2, the cooling degree days, also contains sanitized data. Although this may impact the quality of results to some extent, the pattern of dependency observed here is consistent with engineering studies and suggests a non-linear decrease in natural gas usage as the number of cooling degree days goes up.

Direction 3 shows the dependency between NG intensity and the total square footage of the house. The suggested relationship is linear over the range of square footage where the most observations are concentrated. So the natural gas demand

grows linearly for households between 900 and 6000 sq.ft. Consumption plateaus after 8000 sq.ft.; however, this occurrence should not be given much emphasis as there are very few points in this range.

Direction 4 represents the effect of the electricity price on demand. Electricity is the primary NG substitute in residential buildings. As expected, the correlation is positive. Increases in electricity prices encourage switching to NG as the primary fuel for the household.

Direction 5 illustrates own price effect. As expected, the correlation is negative and NG price increases result in reductions of NG consumption.

Direction 6 contains data on the temperature setting during the day in winter when someone is home. Natural gas intensity in this direction seems to misrepresent the direction of dependency. The mean of regressor 6 corresponds to the temperature setting of 70F. While there is a positive correlation between temperature setting and energy consumption for the range between 55 and 65 degrees, there is no reasonable explanation why natural gas consumption drops for the ranges from 65 to 80, when the opposite should be observed.

The same can be said about the direction 7, which represents the temperature setting during the day in winter when no one is home. The mean for this regressor is 65F. The base temperature for heating is 65F, so thermostats set to the mean temperature would mean no additional heating is required on a 0 HDD. Thus, it is not clear why direction 7 would indicate a drop in the natural gas consumption while the temperature setting is going up. It might be beneficial to replace these two variables with one that would represent the difference between temperature setting when someone is home and temperature setting when someone is not home. The higher the delta, the less energy is consumed while the building is not occupied. There is also an additional factor that leads to misrepresentation of the relationship for this covariate. All temperature settings data is self-reported. In fact, studies have found that persons often report lower-than-actual thermostat settings, even when they know that their settings are being recorded as shown by Lutzenhiser (1993). No actual readings of the thermostat are taken. As saving energy becomes a more widely-publicized topic, respondents understate heating temperature settings, as well as misreport the way programmable thermostats are used, to fall within the range they perceive as socially acceptable. On the other hand, data on natural gas consumption comes directly from the bill and reflects actual consumption levels. Therefore, even

restructuring the variable may not produce a desirable result using existing data.

Direction 8 represents association between the level of natural gas intensity and temperature during the sleeping hours in winter. As the setting goes up from 50 to 70F so does the NG consumption. The slight drop in the gas usage around that point is unexpected. The concern with temperature setting being self-reported is pertinent here as well, as the owners tend to misreport lowering the thermostat settings. So the houses that are set at much higher temperatures, but underreport to be closer in line with culturally-accepted 65-70F level, will drive the result for this average level much higher than what it should be. The estimated natural gas consumption will be inflated for the misreported temperature and underestimated for the higher temperature intervals that would otherwise correspond to that actual heating requirement. This makes the results to the right of the anchor level appear lower than at the average setting, thus erroneously suggesting negative correlation over this interval of temperatures.

The next group of variables, 9 through 22, are categorical unordered variables. The main concerns with including these variables into the smooth backfitting algorithm was the potential violation of the mean-zero assumption for each direction to meet the identification conditions. Nevertheless, the overall results are reasonable.

Direction 9 contains information on the exterior wall construction material. All other things held equal, the change of the wall type variable leads to the expected change in the NG intensity. The lowest NG consumption is shown for stucco, concrete block and stone. By stucco, residents usually refer to either the synthetic cladding that is applied over polystyrene panels, which provide extra insulation, or to cement plaster (lime sand and Portland cement). If installed properly, the latter seals the house, but not as thoroughly as synthetic systems. Concrete block and stone will serve as thermal mass storage, slowing down heat loss. The highest NG consumption is shown for houses with aluminium/vinyl/steel siding or wood shingles. This is consistent not only with the properties of each material and construction methods associated with it, but also with the vintage of the homes that would have these materials installed. In turn, there is a strong correlation between house vintage and quality of wall insulation.

Direction 10, heated garage, produced reasonable results. Category 0 corresponds to the house with no garage. Category 1 represents the houses where there is a garage, but it is not heated. Attached garage provides additional buffer between the heated

part of the house and the environment, thus slowing down heat loss. The results suggest that heating the garage will increase natural gas consumption by up to 14 kBtu. Complete interpretation of this increase also depends on whether garage space is included in the total square footage of the house or not. Also, this regressor is picking up additional effects impacting NG use. Absence of a garage is more typical of older neighborhoods with lower housing prices. They often share similar quality of construction, amount of insulation and level of equipment. Therefore, fairly high NG intensity for houses with no garage is not an unexpected result.

Direction 11 identifies the relationship between the NG intensity and ownership of the house. The result is reasonable as the owned houses have lower energy consumption as compared to rented (the middle) and occupied without payment (the highest). The difference between three categories is around 4kBtu, with delta between the second and the third category being over 1 kBtu. This is consistent with previously documented results of the Caravan Opinion Research Corporation (ORC) surveys. These surveys showed a higher willingness to invest in the energy-saving solutions and high overall concern about the energy efficiency of the residential structure being more typical for the landlords than the renters. There is also a difference in investment decisions associated with primary dwellings versus rentals or additional houses used by relatives or friends without rent payment.

Direction 12 shows the pattern of association between the NG intensity and type of fuel used by burners for cooking on the stove. The peak value is observed for the household equipped with piped natural gas for cooking. There is no difference between using some other fuel (category 0) and bottled propane (category 2). On one hand, these two categories could be combined. On the other hand, residents usually refer to both types of fuel (propane and natural gas) generally as gas, so it is worth keeping for clarification. There is a 4 kBtu reduction if the household is using electricity for cooking burners, which is a reasonable result. This result can also be partially attributed to multicollinearity in data, namely if the household has piped natural gas then it is expected that burners would use NG, but so would the water heaters, clothes dryer and potentially other systems.

Direction 13 contains information on the fuel used by the dryer. NG dryer (category 1) is associated with highest NG consumption. Presence of electrical dryer is associated with lower NG consumption by about 13 kBtu. No-dryer households (category 0) have the lowest level of gas consumption. The overall difference between

households with NG dryers and no dryers reaches approximately 20 kBtu, but that result could be another manifestation of multicollinearity in the data. Households without dryers are more typical for older neighborhoods with lower housing prices, lower construction quality and lower thermal integrity, as well as lower probability of any retrofits and energy-saving solutions being implemented over the duration of occupancy.

Direction 14 shows the dependency between the NG use and the type of secondary heating equipment installed in the house. Typical secondary heating equipment includes central warm-air furnace with ducts (category 1), steam/hot water system with radiators/convectors in each room or pipes in the floor or walls (category 2), built-in floor/wall pipeless furnace (category 3), built-in room heater (category 4) and wood cooking stove used to heat the house (category 5). Cases of no secondary equipment are included as a category with value 0. The result for this category is intuitive as the households with no secondary equipment will have all the heating load provided by the main equipment. Since the RECS microdataset was filtered to keep only observations with piped natural gas, also intuitive is the result that houses equipped with natural gas intake are more likely to use natural gas as their primary heating fuel. Central warm-air furnace with ducts implies more efficient heat delivery system, therefore reduction of the NG consumption for category 1 is also an expected result.

The resulting increase in NG consumption that occurs when the built-in room heaters provide secondary heat is unexpected. This option includes separate in-room heaters burning oil, kerosene or gas, with the first two fuels being a more widely spread option. It is possible that this result is correlated with thermal integrity of the dwelling, as built-in room heaters are more typical for older houses with lower insulation and construction quality.

Direction 15 describes the relationship between NG consumption and the controls installed in the house. There seems to be no difference in NG consumption if there is a programmable (category 1) or non-programmable (category 0) thermostat in the house. These two categories are associated with increased NG demand. The result for category 3 is counterintuitive as it suggests that absence of thermostat is characterized by a significant reduction in NG consumption. Both the direction of change and the magnitude of 16 kBtu are counterintuitive. The explanation might be that absence of thermostat is dictated by warm climate zone and is an indicator of a non-heated dwelling or very little heating is needed. Although the sample was filtered out to

retain only the residential buildings that are heated, houses that are in need of very little heating and may not be equipped with thermostats are included in the sample.

Behavioral information is contained in directions 16 and 17, which deal with programming thermostat to lower temperature for heat setting at night and, correspondingly, when no one is home. The result is counterintuitive as it suggests that programming the thermostat to lower temperature automatically is associated with higher NG use. Neither the direction of change, nor magnitude (3 kBtu) are intuitive.

Direction 17 also produced a counterintuitive pattern. It indicates that the highest NG consumption is for the houses with thermostats preprogrammed to lower setting when no one is home during the day. Then it drops by about 1 kBtu for the houses that have no thermostats, and drops down even further for houses where the temperature is not lowered. For detailed analysis of these two variables more refined data is needed. To separate out the behavioral impact, it is necessary to also account for climate. Thermal integrity of the building usually is strongly correlated with the climate. In turn, in more severe climate conditions, where NG intensities are the highest, the inhabitants are more likely to adjust thermostats up or down from the base setting.

Direction 18 covers the types of fuels used by the primary heating systems. As expected, NG as primary heating fuel (category 1) would result in the highest NG intensity. If the heating degree days data were not sanitized, it would have been possible to approximately identify the climate zone associated with a particular set of observations. There is a strong dependency between the climate zone and choice of fuel for heating that could impact this result. The lowest NG usage is for the houses heated with kerosene or fuel oil. Natural gas consumption for houses that use electricity as primary fuel goes up by 15 kBtu. This could be explained by the fact that some houses with piped natural gas available use electric-source equipment as their primary heating system. The latter use NG for auxiliary heat. Therefore, in this particular case, NG would be used complimentary to electricity. A similar explanation is valid for increase in NG use by 10 kBtu for dwellings using wood and solar energy as a primary heating fuel.

Correlation between the type of heating equipment providing the heat and NG usage is depicted on the graph for direction 19. The lowest NG usage is suggested where portable electric heaters are used to provide most of the heat (category 9). If the heating load can be met with the portable electric heaters, this would indicate that only very little heating is needed and piped NG is used for water heating and

cooking only. Similar explanation is valid for heating stoves burning wood (category 7), portable kerosene heaters (category 10) and cooking stoves used for heating (category 11). The suggestion of highest NG consumption being characteristic of houses with steam/hot water system and radiators/convectors in each room (category 1) is reasonable. High level of NG consumption shown in the graph is expected as this heating system choice impacts natural gas intensity through water-heating requirement, but it is also a manifestation of the climate zone and age/vintage of the house. NG consumption decreases for houses where heat pump is used as a primary equipment, but it is still higher than any other category. This result can also be explained by complimentary use of NG for the auxiliary system that usually turns on as temperatures fall below certain level as electric heat pump becomes less efficient at very cold temperatures. Relatively low NG consumption, according to the regression results, is associated with using central warm-air furnace system with ducts to individual rooms. Considering that this is one of the more efficient heating distribution systems, this is an expected result. Properly designed duct systems have a significant impact on how much heat is lost during delivery. The newest houses have ducts located in the air-conditioned and heated spaces, which results in even more efficient distribution of heat, thus reducing NG intensity. In addition, this is a manifestation of multicollinearity between the house age, quality of construction/insulation and income level of the household.

Results for regressor 20 represents the type of fuel used to heat water for washing or bathing. As expected, if the primary water heating fuel is NG, its consumption is higher than for other fuels. The overall difference is 24 kBtu.

Direction 21 is of a particular interest as it provides some insight on the relationship between the method of how NG is billed and its consumption level. If the household sees the full bill and pays it all, it seems to suggest the lowest result among all categories. Paying the utility bill in full corresponds to category 0. The consumption increases significantly, on the order of 16 kBtu, if all of the payment gets included in rent (category 1) or the household faces only a portion of the total bill for rented dwelling(category 2). This increase could be attributed to differences in willingness to pay for various technology options or invest in energy efficiency between the renters and the owners residing in the house. The result also suggests the difference in NG consumption due to the signal of NG prices not reaching the consumer, or a behavioral difference due to the "paid for" attitude of the consumer that pays a lump sum

irrespective of the actual usage. Such a result is consistent with currently ongoing research on residential energy-efficiency.

Direction 22 picks up the difference in the natural gas intensity due to someone staying at home the whole day verses the house being unoccupied during working hours. The delta of 1.5 kBtu is straightforward to interpret as it shows a clear NG intensity decrease for unoccupied house.

The next set of variables, 23 through 31, were treated as ordered categorical regressors. As suggested by the results, some of them could have been treated as continuous variables. Moreover, several directional regressions show rather smooth change, which may be suggestive of the particular type of a parametric relationship.

Direction 23 characterizes the impact from the number of stories in the building. The lowest NG consumption is for the one-story building, followed by the split level house and two-story structure. The highest level is for the three-story dwellings. As the number of stories increases, the structure design tends to change towards narrower buildings. This leads to a much higher exchange surface, which explains higher NG intensity for buildings in this category. It is necessary to note that all apartment complexes were excluded from the sample. The results cover only single-family detached housing units.

Direction 24 produced rather interesting result. Category 3, where the entire basement is heated during winter shows highest NG consumption. The second highest demand for NG is shown for the houses that have a basement but do not heat any portion of it (category 1). It is followed by the houses where there is a basement and portion of it is heated. This result appears counterintuitive, but may have reasonable explanation. Unheated basements are typical for older houses with unfinished basements. If a portion of it is heated, it is likely that the thermal integrity of the basement has been improved. The difference between these two categories is 2 kBtu. This directional result could be different if the regressor is restructured as a binary versus ordered categorical variable, such that it does not attempt to account for a particular portion of the basement which measurement is not defined. Also, if the retrofit information were available, it would be possible to analyze its correlation with the vintage of the house.

Direction 25 describes the portion of the attic that is warm, and the results are reasonable. It suggests a linear relationship between the fraction of attic that is heated and NG consumption. The difference between a house with no attic versus a

house with an unheated attic is approximately 4 kBtu. Usually no attic implies a flat roof with not much room for insulation. Just the presence of an attic has a favorable effect, as it provides a buffer zone slowing down the heat loss in addition to allowing better insulation. This is followed by the partially heated attic with increase in NG demand by about 8 kBtu. The highest NG consumption is shown for fully heated attic, which would be expected.

Regression results for direction 26, house vintage, are reasonable. There highest NG consumption is shown for category 0 that represents houses built before 1940. NG demand decreases for the houses built in the 1940's by about 10 kBtu, which is followed by the 1950's vintage. There is an increase in the NG consumption of housing built between 1960 and 1969 up from the level shown for 1950 vintage by 5 kBtu, which may be attributable to changes in construction practices. For houses built between 1970 and 1989 the NG consumption decreases by 8 kBTU, which corresponds to improvements in thermal integrity. This trend reverses for dwellings built after 1990, which can be attributed to several factors. First and foremost this is the period when houses with high ceilings gained popularity. In addition, this market trend was accompanied by a shift in the design away from standard rectangular houses to designs with less conventional angles and additional coves. The latter contributes to lower overall energy-efficiency of the house and the effect is reinforced by the ceiling height leading to even more drastic efficiency loss.

Variable 27, which describes the number of thermostats in the house (from 0 to 6), produced a rather interesting result. The drop in the NG consumption between the category with no thermostat and one thermostat by 1 kBtu is reasonable. Then the consumption increases by 17 kBtu for houses with 2 thermostats. The highest level is registered for 3-thermostat houses leading the previous group by about 2 kBtu. This could be explained by the fact that this variable contains redundant information as number of thermostats is linked to the house size. In addition, the number of thermostats might be a representation of inefficient heating system with individual dial in each room in older houses. For each additional thermostat after 3 the consumption drops.

Direction 28 recovers the dependency between number of rooms not heated during the winter and the NG demand. No particular pattern of dependency can be derived from these results. On the surface it would seem likely that this variable should have inverse impact on NG consumption, as more rooms that are unheated in winter would

imply that less NG should be consumed. However, any unheated space that is not zoned appropriately can contribute to the heating load of a house.

Direction 29 analyzes the building shell component heat load contributions by looking at the windows with various glazing and insulating characteristics. The left side of the chart shows the increase in the natural gas consumption across first three categories (single-paned glass, double-paned glass and double-paned glass with low-E coating). This result is somewhat counterintuitive as it would be expected that number of window panes (e.g. single-paned versus double-paned) should be negatively correlated with energy demand, as improved windows have higher energy efficiency. One possible explanation might be the size difference between older single-paned windows and newer double-paned. There is a trend to increase size of windows or incorporate additional windows when retrofits are implemented. Also, newer houses tend to have higher number of windows, which would also increase heat loss and result in the higher NG consumption. In addition, this can also be affected by the climate. Unfortunately the information on window quantity and sizes is not available to test either one of the assertions. Climate information is not included either. NG consumption goes down for categories with triple-pane glass (category 3) and triple-pane glass with low-E coatings (category 4 and 5), which is expected.

Direction 30 describes the relationship between NG consumption and number of people living in the house. The result is reasonable considering that NG demand would likely increase with each consecutive inhabitant. The magnitude of change is also reasonable, as marginal change decreases with each consecutive occupant. Gas consumption drops by 3 kBtu as the number of inhabitants grows from 5 to 7, suggesting that results could plateau out after a certain number of residents representing economies of scale in NG usage – a reasonable result considering that heating requirements would not change with each consecutive inhabitant and natural gas consumption associated with water heating, cooking and dryer use would go up at a smaller rate.

Direction 31 links the income level with the natural gas consumption of the household. It can be concluded that due to the number of categories this variable should be treated as continuous. Initially there is a slight drop in NG intensity as the income grows from less than \$2500 to approximately \$25,000. As income grows, an increase in NG consumption is observed. Categories 11 through 18 correspond to the income interval from \$45,000 to \$85,000. Income at these levels would at least be partially

linked to the type of the house, quality of construction, level of insulation and types of equipment serving the household and this would likely be another representation of the multicollinearity in the data. This increase is followed by a drop in NG consumption for income categories in excess of $85,000. It can be attributed not only to the direct effect due to change in willingness to invest in the energy-efficient solutions, but also a change in level of education, environmental considerations, as well as the shift in the initial quality of occupied homes.

2.6 Conclusion

This essay employs an econometric approach to analyzing natural gas consumption intensity of residential buildings that can be used in combination with simulations for describing the impact of various household and structure attributes on energy demand. The econometric approach employed uses a local linear smooth backfitting estimator, which is extended to include categorical variables. Satisfactory results were obtained for the majority of the covariates and the estimation technique was able to accommodate a correlated set of mixed data.

Nonparametric regression estimation revealed patterns of dependency that could not have been achieved by parametric analysis. Some of the results were suggestive of particular parametric relationships. However, these relationships were only sustained over a portion of the regressor range, as the overall result has the appearance of several superpositioned parametric associations depending on what interval of the regressor support is considered.

This analysis could be extended by combining smooth backfitting regression with stochastic frontier estimation via the method suggested by Fan, Li and Weersink (1996) and, more importantly, by using generalized profile likelihood framework of Severini and Wong (1992). The comparison can be done across residential buildings or groups of residential building based on the ranked efficiency score. The regression portion of the analysis would provide the ability to interpret the efficiency scores from the energy management stand point as a combination of efficiency scores along with each directional regression result allows further investigation of possible causes. This approach could also provide information on the selection of building technologies and engineering and behavioral solutions that could potentially improve the level of energy intensity of residential buildings. One of the issues with using the suggested approach

is to clearly understand how a production frontier can be defined within the context of natural gas usage by residential buildings. If it was possible to isolate only the information that is related to heating, then the thermostat setting could be used as a proxy for the output. The efficiency of maintaining the dwelling at that temperature while all other inputs, attributes and characteristics vary could be compared through ranking. Clusters of houses with similar ranking would provide an insight into what primary features, behavioral characteristics, and house attributes impact the ability to maintain residential buildings at a set temperature.

The benefit of the current analysis is three-fold. The main result, which is the directional impact of each covariate, can be utilized for in-sample prediction to approximate energy demand of a residential building whose characteristics are described by the regressors used in this analysis, but a certain combination of their particular values does not exist in the real world. The only caution is that the best estimates are for the interior of the intervals where the regressors take values. The closer the values are to the end-points of the regressor range, the less accurate are the results.

The second benefit is the information on how natural gas demand might change once a particular characteristic or attribute is altered. For continuous variables the local linear framework applied in this essay produces the values of the slope at each observation as part of the estimation procedure. As far as the categorical variables are concerned, the slope estimates are not calculated as part of the procedure, but they can be easily computed by comparing change in the natural gas usage while moving from one category to another for each of the regressors. For example, results on wall construction material suggest that the natural gas consumption goes down by about 8 kBtu for houses with composite (shingle) siding versus houses with vinyl siding. Properly installed stucco siding may reduce the gas consumption even further (by about 10 kBtu). Jointly with the cost estimates of such improvements this results can be used as a quick tool for benefit-cost analysis of residential upgrades and retrofits under a fixed budget.

The third and the most obvious result follows along the lines of the previously discussed benefit, but with a very particular implication. It shapes the message that changing, for example, the thermostat temperature setting several degrees up or down while holding everything else fixed has a very tangible effect on natural gas usage and related household energy expenditures. Another behavioral result is the relationship between natural gas consumption and billing method. Seeing the full

bill and paying it in full corresponds to the lowest energy consumption level. The consumption increases significantly if a household faces only portion of the bill, or if the full payment is included in rent and the actual consumer never sees either the amount of natural gas consumed, or associated monthly expenditures. The link is obvious, the link is measurable, and the result is produced by a nonparametric estimation procedure without imposing a particular specification on the shape of that relationship.

The primary objective of this analysis was to investigate the applicability of a particular nonparametric methodology to quantifying the impact of behavioral variables using econometric methods. Behavioral aspects of energy usage are largely treated by traditional parametric models as an unobservable effect. If good-quality microdata is available on behavioral aspects of energy usage, it is possible to extend this nonparametric analysis to a larger number of regressors and encompass the relationship between behavioral changes and energy usage at a more refined level.

3 Semiparametric estimation of stochastic production frontier with additivity constraints

3.1 Introduction

This paper analyses the efficiency of fossil-fueled power generation units. The topic of interest is changes in power generation efficiency that could result from the currently debated emission-reduction policies. The objective of this paper is to establish a benchmark for such comparisons by estimating and ranking efficiency scores of existing generation units.

This study is different from previously published research as it contains an application of a semiparametric stochastic frontier model with the introduction of an additivity constraint on the production function of electricity generation. This methodology uses a fully-defined estimator and results in a completely operational procedure for frontier estimation. The method is applicable to the semiparametric setting where the distribution of the error term is specified and the additive separability of the production function is an appropriate assumption. Here we adopt the error term specification of Fan et al. (1996) with an additivity constraint accommodated by following the methodology of Mammen et al. (1999). The asymptotic properties of this combined estimator have not been established in this paper, but the theoretical properties of the estimator defined by Fan et al. (1996) have been derived in Martins-Filho and Yao (2009). The asymptotic behavior of smooth backfitting was established in Mammen et al. (1999).

The additivity in the model proposed here is accommodated using smooth backfitting estimation, which is different from the classical backfitting of Buja et al. (1989). The latter estimator is not efficient according to the "oracle efficiency" criterion put forth by Linton and Nielsen (1995), which means obtaining a directional regression estimator that is asymptotically the same as the estimator when all other directions are known. It was shown by Opsomer and Ruppert (1997) and Opsomer (2000) that backfitting does not reach this oracle efficiency bound. Furthermore, the classical backfitting estimator is not known to be asymptotically normal.

Smooth backfitting is oracle efficient and has the intuitive geometrical interpretation of a projection of the data onto the space of additive functions. Smooth backfitting possesses a high degree of implementational appeal as its iterative equations rely on the estimation of univariate regressions for each covariate, as well as

univariate and bivariate densities only. In addition, smooth backfitting is capable of satisfactorily accommodating covariates with a significant degree of correlation as demonstrated by Nielsen and Sperlich (2005).

This essay is organized as follows. Section 3.2 provides a brief summary of the recent studies analyzing efficiency of power generation, as well as a review of the literature dealing with the econometric estimation of production frontiers. Section 3.3 explains the model under consideration and provides a description of the estimator. It also contains the discussion of bandwidth choice for smooth backfitting estimation of the conditional mean. Section 3.4 presents an efficiency study using the data on energy output, fixed operation and maintenance (O&M) cost, variable O&M cost, and fuel cost of 394 power-generating units in the U.S. A summary of the computational algorithm is included as well. Section 3.5 contains the results and an outline for future research.

3.2 Literature overview

Efficiency of the electricity generating industry has been a focus of multiple studies. Nelson (1984), Baltagi and Griffin (1988) and Callan (1991) analyzed productivity change in the electric utility industry. Emissions were not included as bad outputs, but the cost data reflected the input cost associated with pollution controls. McDonnell (1991) estimated a translog cost model with various fuels using a cross-section of 82 privately owned utilities for the year 1987. Results of this study emphasized high substitutability between gas and coal. Kleit and Terrell (2001) explored the potential production efficiency gains and associated cost reduction for 78 steam plants. Hiebert (2002) investigated the operating cost efficiency of generating plants from 1988 through 1997. The results showed that average operating efficiency increased in those states that were undergoing a transition to retail competition.

The recent research by Pasurka (2003) characterized the relationship between changes in SO_2 emissions, technical efficiency, changes in the output mix and input growth. Their results suggested a strong dependency between the changes in the output mix and changes in SO_2 emissions. Dorfman and Atkinson (2005) analyzed productivity and efficiency in the presence of undesirable inputs. They utilized a parametric approach to estimate the shadow prices, technical efficiency and productivity changes for a panel of 43 privately-owned electric utilities.

Previous literature that deals with efficiency in the context of frontier estimation has evolved in two main directions: deterministic and econometric. The deterministic estimation of frontiers is represented by approaches such as Data Envelopment Analysis based on the work of Farrell (1957) with extensions to the method described in Seiford and Thrall (1990), and Free Disposable Hull defined by Deprins et al. (1984). The econometric approach was first introduced by Aigner et al. (1977), Meeusen and van den Broeck (1977), and Battese and Corra (1977), and it relies on a full parametric specification of the production function, as well as the probability density function of the error term. Although the econometric approach allowed incorporating a stochastic component into the model, the assumptions on the production function and error term were extremely limiting. This work was later continued by Greene (1990), who extended the model by relaxing the assumption of the one-sided disturbance being distributed as truncated normal and using the more flexible Gamma distribution. Fan et al. (1996) removed the parametric restrictions on the production function, but kept the same distributional assumptions on the error term components as Aigner et al. (1977). Fan et al. (1996) assumed that a one-sided error, representing technical inefficiency, was an independent identically-distributed (IID) normal random variable, truncated at zero. The component that represented statistical noise was assumed to be a two-sided IID normally distributed error. Although a semiparametric estimation methodology was developed and the resulting estimator fully defined, no asymptotic properties were established. Some of the recent frontier estimation studies such Henderson and Simar (2005) and Kumbhakar et al. (2007) have proposed fully nonparametric stochastic frontier techniques based on local linear least squares regression and local maximum likelihood. The first study relied on panel data; the second study used cross-sectional data. The most recent analysis by Kumbhakar et al. (2009) used the two-step procedure to jointly analyze technology choice and technical efficiency with application to organic and conventional farming.

This essay addresses the topic first discussed in this section - efficiency and productivity of the power generation across several types of units. The approach utilized here is econometric frontier estimation, thus the paper takes the efficiency analysis in the second direction mentioned above. The semiparametric model of Fan et al. (1996) is used as a foundation for this analysis, but our model takes it a step further by incorporating an additivity restriction on the production function. This is handled nonparametrically, relying on the smooth backfitting framework of Mammen

et al. (1999). The inclusion of categorical variables to account for the differences in the generation cycle was motivated by the technology choice discussed jointly with technical efficiency in the latest study by Kumbhakar et al. (2009). Usually this is handled parametrically or using the frequency approach. The procedure employed here is different as it includes kernel smoothing of the categorical variables inside of the smooth backfitting estimation. To summarize, this paper extends the work of Fan et al. (1996) to incorporate some of the advances in nonparametric estimation and results in a fully operational estimation procedure with a straight-forward algorithm, which is capable of nonparametrically handling categorical variables to model the production function, and applies this estimator to estimating efficiency of power generation, the topic that has been widely discussed in the efficiency literature since the 1960's.

3.3 Model

This section contains the description of the semiparametric model and an overview of the proposed estimation approach. The semiparametric stochastic production frontier model considered in this analysis is similar to Fan et al. (1996), but the production function is restricted to be additively separable, namely:

$y_i = g(x_i) + \varepsilon_i$, $\varepsilon_i = \nu_i - u_i$,

where u_i is a one-sided error term representing technical inefficiency distributed as a normal truncated at zero with expected value μ_u and variance σ_u^2, and ν_i is a two-sided error term distributed as normal IID with zero mean and variance σ_v^2. The single output production function here is denoted as $y \leq g(x)$, where x is the vector of inputs. Further, we impose additivity on the production function, i.e. $g(x_i) = m_0 + \sum_{j=1}^{d} m_j(x_j)$. The stochastic frontier for this model is defined as $SF_i = g(x_i) + \nu_i$. Random external shocks that influence efficiency of the production unit, but are outside its control, are represented by ν_i. Variable u_i captures technical inefficiency due to the factors that are within the control of a production unit. It allows a firm to be inefficient relative to its maximum possible level of output given by its stochastic frontier, which already accounts for the effect of external random factors outside the firm's control.

The distribution of the composite error term, derived as the sum of a normal and

truncated normal random variables, was first given in Weinstein (1964):

$$f_{\varepsilon_i}(\varepsilon_i) = \frac{2}{\sqrt{2\pi\left(\sigma_u^2+\sigma_v^2\right)}}\exp\left(-\frac{1}{2}\frac{\varepsilon_i^2}{\left(\sigma_u^2+\sigma_v^2\right)}\right)\left[1-\int_{-\infty}^{\varepsilon_i\frac{\sigma_u}{\sigma_v\left(\sigma_u^2+\sigma_v^2\right)^{1/2}}}\phi\left(\tau\right)d\tau\right]$$
$$= \frac{2}{\sigma}\phi\left(\frac{\varepsilon_i}{\sigma}\right)\left[1-\Phi\left(\frac{\varepsilon_i\lambda}{\sigma}\right)\right],$$

where $\lambda = \frac{\sigma_u}{\sigma_v}$, $\sigma^2 = \sigma_u^2+\sigma_v^2$, $\phi\left(\cdot\right)$ −standard normal density function and $\Phi\left(\cdot\right)$ is a standard normal distribution function. After the appropriate transformation, the conditional density of y is defined as

$$f_{y_1\ldots y_n|x}\left(y_1\ldots y_n|x\right) = \left(\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\right)^n\exp\left[-\frac{\sum_{i=1}^n\left(y_i-g(x_i)\right)^2}{2\sigma^2}\right]$$
$$\times\prod_{i=1}^n\left[1-\Phi\left(\left(y_i-g(x_i)\right)\frac{\lambda}{\sigma}\right)\right].$$

The natural choice of an estimation procedure is maximum likelihood. The conditional log-likelihood function is

$$\begin{aligned}\tilde{L} &= \ln f_{y_1\ldots y_n|x}\left(y_1\ldots y_n|x\right)\\ &= \frac{n}{2}\ln\left(\frac{2}{\pi}\right)-\frac{n}{2}\ln\sigma^2-\frac{\sum_{i=1}^n\left(y_i-g\left(x_i\right)\right)^2}{2\sigma^2}+\sum_{i=1}^n\ln\left[1-\Phi\left(\frac{\left(y_i-g\left(x_i\right)\right)\lambda}{\sigma}\right)\right]\end{aligned}$$

with unknown components $g\left(x_i\right)$, λ and σ^2. Assuming that the solutions are on the interior, first order conditions are derived by taking the derivatives with respect to the unknown parameters and setting them equal to zero. Due to the non-linearity of the first order conditions (FOC), a closed form solution cannot be obtained. Since the unknown $g\left(x_i\right)$ is present in the FOC, direct estimation of the parameters σ^2 and λ is not operational, even if a concentrated maximum likelihood estimation is attempted.

Since the conditional expectation of y_i given x_i cannot be separated from $g\left(x_i\right)$, it is suggested to replace this function in the FOC by $g\left(x_i\right) = E(y_i|x_i)+\mu_u$, where

$\mu_u = \frac{\sqrt{2}}{\sqrt{\pi}}\sigma_u = \frac{\sqrt{2}\sigma\lambda}{\sqrt{\pi\left(1+\lambda^2\right)}}$. The resulting FOC is of the form

$$\frac{\partial \tilde{L}}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}\left(y_i - E(y_i|x_i) - \mu_u\right)^2}{2\sigma^4}$$
$$+ \frac{\lambda}{2\sigma^3}\sum_{i=1}^{n}\left[\frac{\phi(\lambda\left(y_i - E(y_i|x_i) - \mu_u\right)/\sigma)}{1-\Phi(\lambda\left(y_i - E(y_i|x_i) - \mu_u\right)/\sigma)}\right]\left(y_i - E(y_i|x_i) - \mu_u\right) = 0,$$

$$\frac{\partial \tilde{L}}{\partial \lambda} = \frac{1}{\sigma}\sum_{i=1}^{n}\left[\frac{\phi(\sigma\left(y_i - E(y_i|x_i) - \mu_u\right)/\sigma)}{1-\Phi(\lambda\left(y_i - E(y_i|x_i) - \mu_u\right)/\sigma)}\right]\left(y_i - E(y_i|x_i) - \mu_u\right) = 0.$$

Using the fact that the second term in the first FOC equals zero, Fan et al. (1996) obtained $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(E(y_i|x_i) - \mu_u\right)\right)^2$. Substituting $\mu_u = \frac{\sqrt{2}\sigma\lambda}{\sqrt{\pi\left(1+\lambda^2\right)}}$ resulted in a quadratic equation $ax^2 + bx + c = 0$, where $a = -\left[1 - \frac{2\lambda^2}{\pi\left(1+\lambda^2\right)}\right]$, $b = -\frac{2\lambda\sqrt{2}}{\sqrt{\pi\left(1+\lambda^2\right)}}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - E(y_i|x_i)\right)$, and $c = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - E(y_i|x_i)\right)^2$. It was suggested that since $b = O_p\left(n^{-1/2}\right)$, the parameter σ^2 can be estimated by

$$\tilde{\sigma}^2 = \frac{c}{|a|} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - E(y_i|x_i)\right)^2}{\sqrt{1 - \frac{2\lambda^2}{\pi\left(1+\lambda^2\right)}}}.$$

As suggested in Fan et al. (1996) the estimation of $E(y_i|x_i)$ is done using a nonparametric regression. The additivity constraint on the production function allows utilizing the smooth backfitting approach of Mammen et al. (1999). The regression model considered here for estimating the conditional expectation of y_i is of the form $E(y|x_1,...,x_d) = m_0 + \sum_{j=1}^{d} m_j\left(x_j\right)$, where $(y, x_1...,x_d)$ is a random vector in $\mathbb{R}^{d+1}$ and we assume that there is a random sample $\{y_i, x_{i1}...,x_{id}\}_{i=1}^{n}$ of $(y, x_1...,x_d)$, m_0 is an unknown scalar parameter, $m_j\left(x_j\right)$ is a sufficiently smooth function for all j, and θ_j is the first order derivative of $m_j\left(x_j\right)$. Also for identification purposes, $E\left(m_j\left(x_j\right)\right) = 0$.

Let $K_h\left(x_{ij} - x_j\right) = \frac{1}{h}K\left(\frac{x_{ij}-x_j}{h}\right)$ be a kernel function such that $\int K\left(\phi\right)d\phi = 1$, and $\int \phi K\left(\phi\right)d\phi = 0$. $h = h(n)$ is a bandwidth such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$, and conditions B(1), B(2')-B(4') of Mammen et al. (1999) are met. For categorical variables, the kernel shape suggested by Aitchison and Aitken (1976) for

the distribution estimation

$$L(x_{it}, x_t, \lambda_t) = \begin{cases} 1-\lambda_t, & \text{if } x_{it} = x_t \\ \lambda_t/\left(c_t - 1\right), & \text{if } x_{it} \neq x_t \end{cases} \quad t = 1, ..., T$$

is used for unordered categorical regressor. The overview of the estimator here relies on the discussion in the first essay.

The multivariate discrete data kernel is defined as $\prod_{t=1}^{T} L(x_{it}, x_t, \lambda_t)$, with joint density of discrete variables being estimated by $\widehat{p}(x_{_1}, ..., x_d) = n^{-1} \sum_{i=1}^{n} \prod_{t=1}^{T} L(x_{it}, x_t, \lambda_t)$. The multivariate kernel for mixed data is

$$W\left(x_{ij}, x_j, h, x_{it}, x_t, \lambda_t\right) = \sum_{i=1}^{n} \prod_{j=1}^{d} K_h\left(x_{ij} - x_j\right) \prod_{t=1}^{T} L(x_{it}, x_t, \lambda_t).$$

The local linear smooth backfitting estimator for mixed continuous and categorical data is a projection of the local linear estimator for mixed regressors onto the space of additive functions. The mixed data local linear smooth backfitting estimator $\widetilde{m}^*(x)$ is defined as the argument that minimizes the following objective function

$$\int \sum_{i=1}^{n} \left[y_i - m_0 - \sum_{j=1}^{d} m_j(x_j) - \sum_{t=1}^{T} m_t(x_t) - \sum_{j=1}^{d} \theta_j\left(x_{ij} - x_j\right) \right]^2$$
$$\times \prod_{j=1}^{d} K_h\left(x_{ij} - x_j\right) \prod_{t=1}^{T} L(x_{it}, x_t, \lambda_t) dx,$$

The minimization is done with respect to $m_0, m_1, ..., m_d$, and all first derivatives $\theta_j(x_j)$. The marginal and bivariate densities here are defined in the same manner as shown in the first essay. For continuous regressors let

$$\begin{aligned} \widehat{p}_j(x_j) &= n^{-1} \sum_{i=1}^{n} K_h\left(x_{ij} - x_j\right), \quad \widehat{p}_j^{j}(x_j) = n^{-1} \sum_{i=1}^{n} K_h\left(x_{ij} - x_j\right)\left(x_{ij} - x_j\right), \\ \widehat{p}_j^{jj}(x_j) &= n^{-1} \sum_{i=1}^{n} K_h\left(x_{ij} - x_j\right)\left(x_{ij} - x_j\right)\left(x_{ij} - x_j\right), \end{aligned}$$

$$\widehat{p}_{jk}(x_j,x_k) = n^{-1}\sum_{i=1}^{n} K_h\left(x_{ij}-x_j\right) K_h\left(x_{ik}-x_k\right),$$
$$\widehat{p}^{k}_{jk}(x_j,x_k) = n^{-1}\sum_{i=1}^{n} K_h\left(x_{ij}-x_j\right) K_h\left(x_{ik}-x_k\right)\left(x_{ik}-x_k\right),$$
$$\widehat{p}^{jk}_{jk}(x_j,x_k) = n^{-1}\sum_{i=1}^{n} K_h\left(x_{ij}-x_j\right) K_h\left(x_{ik}-x_k\right)\left(x_{ij}-x_j\right)\left(x_{ik}-x_k\right),$$

For marginal densities of categorical variables and joint densities for mixed data let

$$\widehat{p}_{t}(x_t) = n^{-1}\sum_{i=1}^{n} L\ (x_{it},x_t,\lambda_t),$$
$$\widehat{p}_{jt}(x_j,x_t) = n^{-1}\sum_{i=1}^{n} K_h\left(x_{ij}-x_j\right)\ L\ (x_{it},x_t,\lambda_t),$$
$$\widehat{p}^{j}_{jt}(x_j,x_t) = n^{-1}\sum_{i=1}^{n} K_h\left(x_{ij}-x_j\right)\ L\ (x_{it},x_t,\lambda_t)\left(x_{ij}-x_j\right)$$

Using similar notation as in the first essay

$$\begin{aligned}
\widetilde{m^*_j}(x_j) &= A - \sum_{t\neq j}^{T}\int \widetilde{m_t}(x_t)\frac{\widehat{p}_{jt}(x_j,x_t)}{\widehat{p}_j(x_j)}dx_t - \widetilde{\theta^*_j}(x_j)C \\
&= A^* - \widetilde{\theta^*_j}(x_j)C, \\
\widetilde{m^*_j}(x_j) &= B - \sum_{t\neq j}^{T}\int \widetilde{m_t}(x_t)\frac{\widehat{p}^{j}_{jt}(x_j,x_t)}{\widehat{p}^{j}_j(x_j)}dx_t - \widetilde{\theta^*_j}(x_j)D \\
&= B^* - \widetilde{\theta^*_j}(x_j)D, \\
\widetilde{\theta^*_j}(x_j) &= \frac{A^*-B^*}{C-D}.
\end{aligned}$$

The smooth backfitting estimates of $\widetilde{m_0}$, $\widetilde{m_j}$ and $\widetilde{\theta_j}$ for continuous variables are obtained by iteratively solving the two equations below for each regressor $j = 1, ...d$

$$\widetilde{m^*_j}(x_j) = A^* - \widetilde{\theta^*_j}(x_j)C, \qquad \widetilde{\theta^*_j}(x_j) = \frac{A^*-B^*}{C-D}$$

The iterative equation for discrete regressors x_t, $t = 1, ..., T$ is

$$\begin{aligned}\widetilde{m_t^*}(x_t) &= \frac{n^{-1}\sum_{i=1}^{n} L\ (x_{it}, x_t, \lambda_t) y_i}{\widehat{p}_t(x_t)} - \sum_{j=1}^{d}\int \widetilde{m_j}(x_j)\frac{\widehat{p}_{jt}(x_j, x_t)}{\widehat{p}_t(x_t)}dx_j - \widetilde{m_0}(x) \\ &\quad - \sum_{k\neq t}^{T}\int \widetilde{m_k}(x_k)\frac{\widehat{p}_{kt}(x_k, x_t)}{\widehat{p}_t(x_t)}dx_k - \sum_{j=1}^{d}\int \widetilde{\theta}_j(x_j)\frac{\widehat{p}^{j}_{jt}(x_j, x_t)}{\widehat{p}_t(x_t)}dx_j.\end{aligned}$$

where $\widetilde{m_0} = n^{-1}\sum_{i=1}^{n} Y_i$.

Once the estimates are obtained for the intercept and each direction, the conditional mean is estimated as $\widetilde{E}(y_i|x_i) = \widetilde{m_0} + \sum_{j=1}^{d}\widetilde{m_j}(x_j)$. The estimator for σ^2 is defined as $\widetilde{\sigma}^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widetilde{E}(y_i|x_i)\right)^2}{\sqrt{1-\frac{2\lambda^2}{\pi(1+\lambda^2)}}}$, which is then substituted into the original log-likelihood function $\tilde{L}$. The concentrated log-likelihood is maximized with respect to the single parameter λ. Then the frontier production function is estimated as $\widetilde{g}(x_i) = \widetilde{E}(y_i|x_i) + \widetilde{\mu} = \widetilde{E}(y_i|x_i) + \frac{\sqrt{2}\widetilde{\sigma}\widetilde{\lambda}}{\sqrt{\pi\left(1+\widetilde{\lambda}^2\right)}}$, which means shifting up the "average" production function by the expected value of the inefficiency term. The individual expected inefficiency score for a particular observation is calculated in the same manner as in Jondrow et al. (1982), i.e.

$$\widetilde{E}(u|\varepsilon) = \left(\frac{\widetilde{\sigma}_u^2\widetilde{\sigma}_v^2}{\widetilde{\sigma}^2}\right)^{1/2}\left[\frac{\phi\left(\widetilde{\varepsilon}\widetilde{\lambda}/\widetilde{\sigma}\right)}{1-\Phi\left(\widetilde{\varepsilon}\widetilde{\lambda}/\widetilde{\sigma}\right)} - \frac{\widetilde{\varepsilon}\widetilde{\lambda}}{\widetilde{\sigma}}\right],$$

where estimates $\widetilde{\sigma}_u$ and $\widetilde{\sigma}_v$ are obtained from $\widetilde{\lambda} = \widetilde{\sigma}_u/\widetilde{\sigma}_v$ and $\widetilde{\sigma}^2 = \widetilde{\sigma}_u^2 + \widetilde{\sigma}_v^2$. In this context, the firm-specific inefficiency score $\widetilde{E}(u|\varepsilon)$ represents how far the firm is operating below its own frontier due to the specific factors that lie within the firm's control. The technical efficiency index can be estimated as $\widetilde{TEI} = y_i/\widetilde{SF}_i$, where $\widetilde{SF}_i$ estimates a firm-specific stochastic frontier. The latter is defined as $SF_i = g(x_i) + \nu_i$.

3.3.1 Bandwidth selection

Several different methods for selecting bandwidths for SBE estimation were analyzed recently. Mammen and Park (2005) introduced a bandwidth selection method

for smooth backfitting based on minimizing the penalized sum of squared residuals. They also compared two additional plug-in methods for the local linear SBE. Moreover, it was suggested that the penalized sum of squared residuals was asymptotically equivalent to cross-validation since this holds true for the classical nonparametric regression as in Härdle et al. (1988).

Leave-one-out least squares cross-validation is recommended for bandwidth selection by Nielsen and Sperlich (2005). It has a particular implementational advantage for local linear smooth backfitting if the underlying relationship is additive, since in this case the SB estimator has additively separable bias and variance. Bandwidth selection is based on minimizing the mean-integrated squared error $MSE(h_1, ...h_d, \lambda_1, ...\lambda_d) = \int E\left[\widetilde{m}(x) - m(x)\right]^2 p(x)dx$. Due to the separability of bias and variance, the mean-integrated squared error for overall regression can be defined as

$$MSE(h_1, ...h_d, \lambda_1, ...\lambda_d) = \sum_{j=1}^{d+T} MSE_j(x_j),$$

where $MSE_j(x_j)$ is the mean-integrated squared error for each regression direction $m_j(x_j)$. Thus the cross-validation problem of minimizing $CV = \sum_{i=1}^{n} \left[y_i - \widetilde{m}^{-i}(x)\right]^2$, where $\widetilde{m}^{-i}(x)$ is the leave-one-out estimator with observation (y_i, x_i) left out of the computation, can be separated. It reduces to performing optimal bandwidth search for each directional regression sequentially. Nielsen and Sperlich (2005) suggest taking starting bandwidths $h_1, ..., h_d$ that undersmooth for each direction and run initial SBE estimation. Then while using one-dimensional grid search to minimize the cross-validation criteria with respect to h_j only, the bandwidth for direction j, bandwidths for all other directions are kept at their starting values. This is repeated for each direction j individually. It is noted that not only it is unnecessary to use leave-one-out estimators for all other directions $m_k(x_k)$, $k \neq j$, while searching for optimal bandwidth for estimation of $m_j(x_j)$, but also all $\widetilde{m_k}(x_k)$ do not need to be estimated at their optimal bandwidth. As shown by Mammen and Park (2005), this procedure results in bandwidths that are optimal for the estimation of the overall regression. If the primary focus of the estimation is accuracy of each single additive component, they suggest using plug-in bandwidths that minimize average weighted squared error

(ASE) for each direction defined as

$$ASE_j(x_j) = n^{-1} \sum_{i=1}^{n} w_j^{-i}(x_j) \left[\widetilde{m}_j(x_j) - \widetilde{m}_j^{-i}(x_j)\right]^2,$$

where $\widetilde{m}_j^{-i}(x_j)$ is the leave-one-out estimator of $m_j(x_j)$ and w_j is a weight function.

This paper adopts a simpler method for bandwidth selection. Since smooth backfitting requires computing the unrestricted regression estimates, as well as univariate and bivariate densities for continuous and categorical data, we use four different bandwidth selection routines. To estimate densities for categorical variables we use the cross-validation method of Li and Racine (2007), where the bandwidth λ is chosen separately for each regressor to minimize

$$CV_p(\lambda) = \sum_{x_c \in S_c} \left[\widehat{p}\,(x_c)\right]^2 - 2n^{-2} \sum_{i=1}^{n} \sum_{v \neq i}^{n} L_{\lambda,iv}$$

where $L_{\lambda,iv}$ is the previously defined kernel with observation $v = i$ excluded from the computation, $S_c = \{0, ...c_t - 1\}$ is the support of x_c and c is the category index. For unrestricted regression estimation for categorical variables, the cross-validation of Li and Racine (2007) is employed. Bandwidth is chosen to minimize

$$CV_{reg}(\lambda) = n^{-1} \sum_{i=1}^{n} \left[y_i - \widehat{m_j}^{-i}(x_j)\right]^2$$

for each j, where $\widehat{m_j}^{-i}(x_j)$ is the leave-one-out Nadaraya-Watson estimator of $m_j(x_j)$ defined as $\widehat{m_j}^{-i}(x_j) = \frac{\sum_{v \neq i}^{n} y_v \, L_{\lambda,iv}}{\sum_{v \neq i}^{n} L_{\lambda,iv}}$. For continuous variables the rule-of-thumb bandwidth selection was used both for estimation of unrestricted univariate regression, as well as densities. Namely, the bandwidth for regression estimation was selected as

$$h_j^{reg} = n^{-1/5} \left\{ s^2 2\sqrt{\pi} \left(\max(x_j) - \min(x_j)\right) \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \left(\widehat{b_3} + \widehat{b_4} x_j + 0.5\widehat{b_5} x_j^2\right)^2\right]^{-1} \right\}^{1/5},$$

where b_3, b_4 and b_5 are estimates of coefficients in regressing the dependent variable y on $\beta_1 + \beta_2 x_j + \beta_3(0.5x_j^2) + \beta_4(\frac{1}{6}x_j^3) + \beta_5(\frac{1}{24}x_j^4)$, and s^2 is estimated in the usual manner based on the residual estimates of this regression. The bandwidth for density

estimation was computed as $hdens_j = (n^{-1/5}) \cdot 1.01a\,(2\sqrt{\pi})^{-1/5}$, and $a = q_{75}(x_j) - q_{25}(x_j)$, where q_{75} and q_{25} are upper and lower quartiles of x_j correspondingly.

3.3.2 Estimation Algorithm

The estimation procedure can be subdivided into two major parts. The first part has to deal with the estimation of the conditional mean via smooth backfitting. Once the estimates of the intercept $\widetilde{m_0}$, directional regressions $\widetilde{m_j}(x_j)$ for each $j = 1, ...d$, as well as first derivatives $\widetilde{\theta}_j(x_j)$ are obtained, the first two components are passed into the second part of the procedure that estimates $\widetilde{\sigma}^2$ and substitutes it into the likelihood function. At this point the numerical optimization is applied in GAUSS to optimize the likelihood with respect to the remaining parameter λ.

The step-by-step algorithm for computation is as follows:

1. Compute the rule of thumb bandwidth for all regression and density estimation as suggested in the previous section.
2. Compute the univariate $\widehat{p}_j(x_j), \widehat{p}^j_j(x_j),\ \widehat{p}^{jj}_j(x_j)$ for all regressors $x_j\ j = 1, ...d$.
3. Compute bivariate expressions $\widehat{p}_{jk}(x_j, x_k), \widehat{p}^k_{jk}(x_j, x_k), \widehat{p}^j_{jk}(x_j, x_k)$ and $\widehat{p}^{jk}_{jk}(x_j, x_k)$.
4. Compute univariate unrestricted estimates for pairs $(\widehat{m}_j(x_j), \widehat{\theta}_j(x_j))$ for each regressor. Save the results as variables m_{old} and θ_{old}.
5. Set the number of smooth backfitting iteration $iter$ to 1.
   (a) For $j = 1$ compute expressions A, B, C, D. Obtain $\widetilde{m_j}(x_j)$ and $\widetilde{\theta}_j(x_j)$, save as m_{new} and θ_{new}. Repeat this step for the rest of continuous variables $j = 2, ...d$. To compute expressions A and B, use updated values from m_{new} and θ_{new} for $k < j$. If $k > j$, use corresponding values from m_{old} and θ_{old}.
   (b) Define a convergence criteria for all j as $\frac{\sum_{i=1}^{n}\left[\widetilde{m_j^{new}}(x_j) - \widetilde{m_j^{old}}(x_j)\right]^2}{\sum_{i=1}^{n}\left[\widetilde{m_j^{old}}(x_j)\right]^2 + \epsilon} < \epsilon$.
6. Set $m_{old} = m_{new}$ and $\theta_{old} = \theta_{new}$. Set $iter = iter + 1$, go to step 5a. Repeat steps 5 and 6 until the convergence criteria is met.

7. Estimate the conditional expectation as $\widetilde{E}(y_i|x_i) = \widetilde{m_0} + \sum_{j=1}^{d} \widetilde{m_j}(x_j)$.

8. Substitute $\widetilde{\sigma}^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widetilde{E}(y_i|x_i)\right)^2}{\sqrt{1-\frac{2\lambda^2}{\pi(1+\lambda^2)}}}$ into the original log-likelihood and maximize it with respect to λ. Since the closed form solution is not available, the numerical optimization should be used.

9. Compute production function estimates, efficiency scores and technical inefficiency index as explained above.

3.4 Results and analysis

This research examines power generation from the standpoint of efficiency of existing generating capacity, which is operated based on the least-cost provision. Emissions are not treated as a bad output, but rather pollution control cost is included as part of the input cost. Data for this study includes fixed operation and maintenance (O&M) cost, variable O&M cost, fuel cost and type of the cycle for 394 generating units as covariates. This information comes from the PROMOD, the electric market simulation package that contains 2008 power generation database.

There are three major categories of generators considered in this study: combined cycle, combustion turbine and steam turbine. Combined cycle is represented by 143 units. There are 115 observations for the combustion turbines and 136 data points for the steam turbines. The last category is subdivided into three subsections to reflect the differences in the fuel used. The following number of observations fall into each subgroup: coal – 73, gas – 24, other – 39. Thus one categorical variable with seven groups reflects the difference in the fuel and cycle type. Fuel cost is used as a proxy for fuel quantity under the assumption that within any category all units face the same input prices, while the differences between categories are captured by the categorical regressor. Output is defined as MWh of unit generation in 2008, which ranges anywhere from 1000 MWh to 15 million MWh.

Before introducing the results of the estimation, it is informative to plot the initial inputs against the output. Figure 1 contains the graph of electricity output against fixed O&M cost. Figure 2 suggests a linear relationship between variable costs and output, as would be expected. Figure 3 shows electricity generation plotted against

the fuel cost. Figure 4 shows the output ranges for each of the categories, which account for the cycle type and fuel type. Category 0 is combined cycle, and category 1 represents gas combustion turbines. The last three groups are all steam turbines, but they use different fuels, thus category 2 is coal-fired, category 3 uses natural gas, category 4 runs on "other" fuel, which includes biomass and heavy fuel oil.

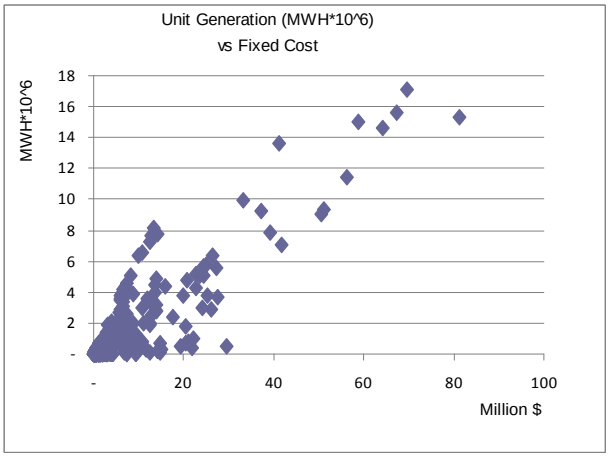

Figure 1. Electricity output versus fixed O&M cost.

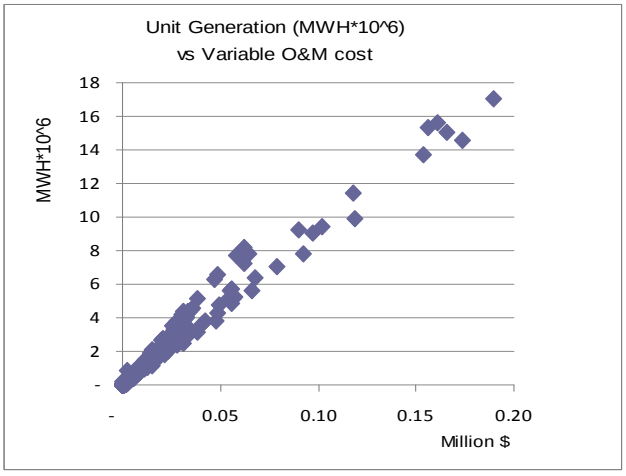

Figure 2. Electricity output versus variable O&M cost

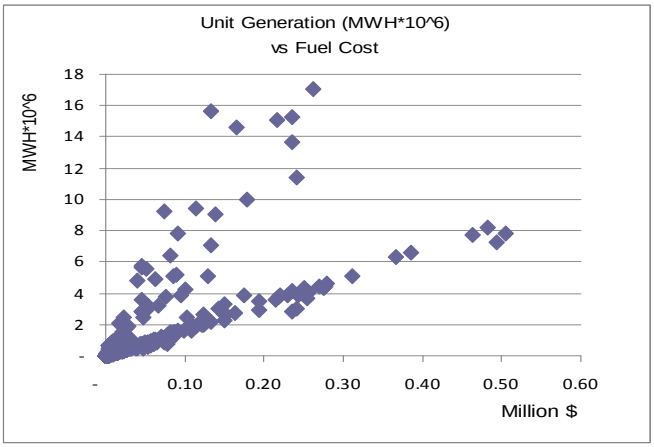

Figure 3. Electricity output versus fuel cost

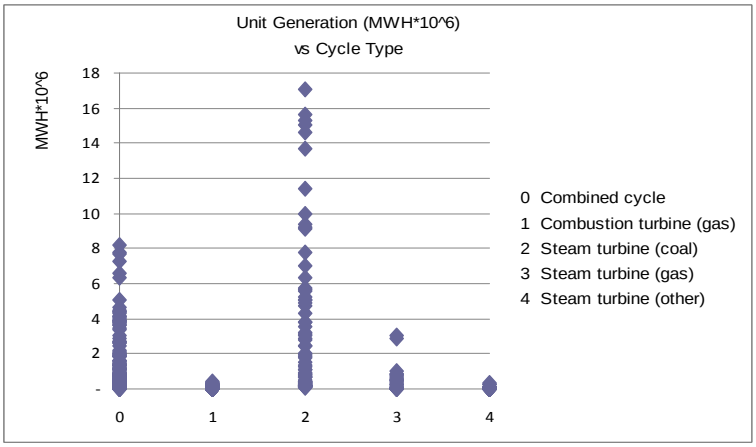

Figure 4. Electricity output versus cycle type

The first part of the estimation, the smooth backfitting procedure, requires bandwidth selection to compute univariate and bivariate densities, as well as unrestricted regression estimates. The bandwidth values for density estimation are 1341.57, 2440.69, 11878.97 and 0.025 for each of the covariates. The regression estimation bandwidth values are 4055.89, 10587.26, 44993.5 and 0.0184 correspondingly.

First the overall result is included for the whole sample, then the estimates are broken down by the turbine and fuel type. This is done with the understanding that

even though the overall estimation is performed for the whole sample, the efficiency should be compared in the context of individual frontiers corresponding to each of the categories. Each of the continuous variables have two plots included. The first plot contains directional estimates for all points of evaluation. The second plot shows a subset of the first plot, but zooms in on the bottom left quadrant of the larger plot to better illustrate the relationship for moderate levels of production away from the boundary region. It is worth noting that the estimation on the boundaries of the plot is less accurate than on its interior, as the closer the observations are to the endpoints, the fewer observations are contained on the boundary side for weighting. Therefore, the patterns of dependency appearing at the edges of the evaluation sample generally lack credibility and should not be given emphasis in the analysis. Note that negative values appear on the Y axis because only the directional regression estimates $\widetilde{m_j}(x_j)$ are presented without the adjustment for the scalar parameter $\widetilde{m_0}$.

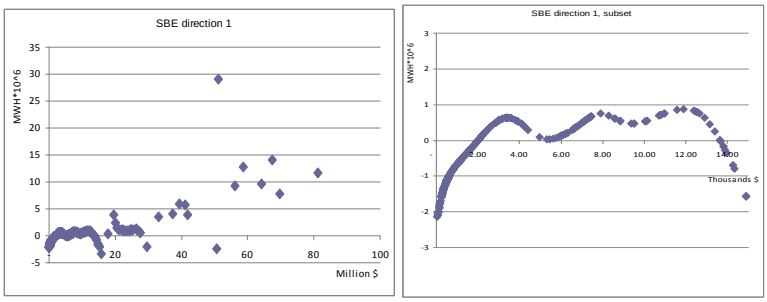

Figure 5. Smooth backfitting estimates for direction 1, fixed O&M cost

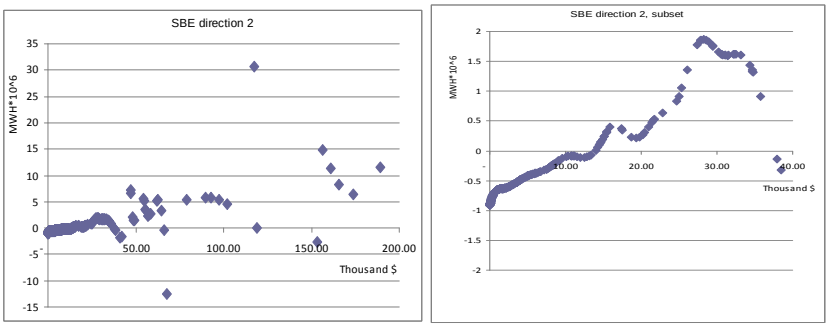

Figure 6. Smooth backfitting estimates for direction 2, variable O&M cost

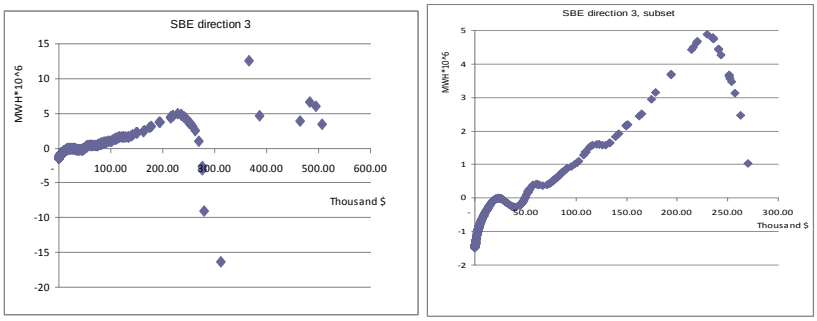

Figure 7. Smooth backfitting estimates for direction 3, fuel cost

Figure 8. Smooth backfitting estimates for cycle type

Once the estimates are broken down by fuel and turbine type, the plots more closely resemble traditionally obtained frontier estimates. The first group of graphs, Figure 9, corresponds to the combined generation cycle. Figure 10 is for the second group, which is combustion turbine. The third group has three subgroups to reflect differences in the fuel type. They are included in Figures 11.1 - 11.3.

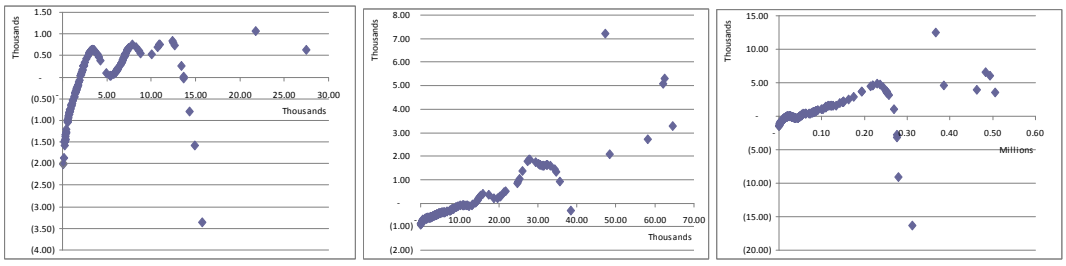

Figure 9. Smooth backfitting estimates for combined cycle

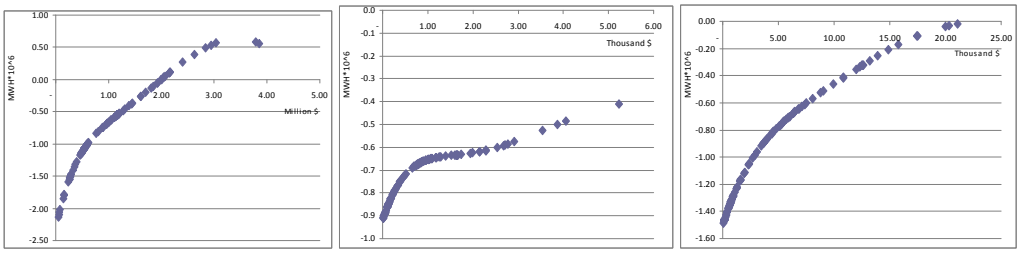

Figure 10. Smooth backfitting estimates for combustion turbine

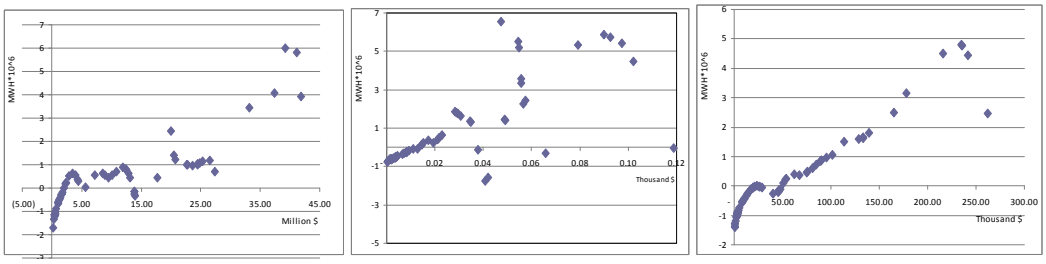

Figure 11.1. Smooth backfitting estimates for steam turbine, coal

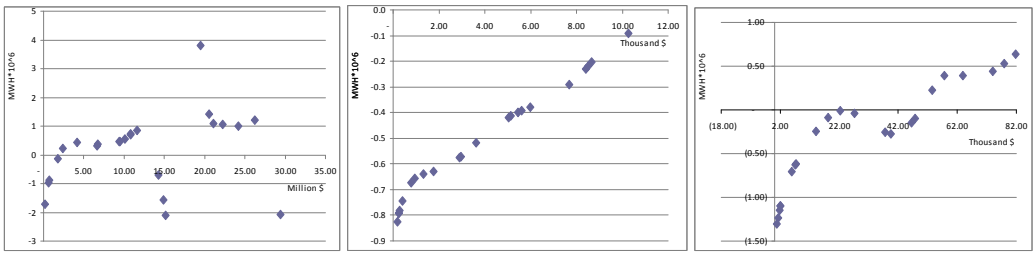

Figure 11.2. Smooth backfitting estimates for steam turbine, gas

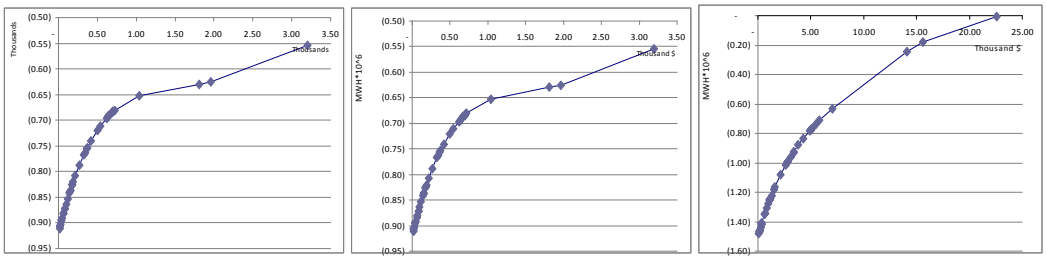

Figure 11.3. Smooth backfitting estimates for steam turbine, other

Based on the smooth backfitting estimates and the expression for variance shown in step 8 of the computational algorithm, the log-likelihood maximization via CML package in GAUSS 9.0 returned 2.0792 as the estimate for λ. Optimization was performed several times using different starting points. The estimate that produced the highest likelihood function value was selected as the final result. Estimate $\widetilde{\sigma}^2$ equals 33,124,692. Estimates for $\widetilde{\sigma}_u^2$ and $\widetilde{\sigma}_v^2$ are obtained from $\widetilde{\lambda} = \widetilde{\sigma}_u/\widetilde{\sigma}_v$ and $\widetilde{\sigma}^2 = \widetilde{\sigma}_u^2 + \widetilde{\sigma}_v^2$ and equal 26,901,863 and 6,222,829.5 correspondingly.

Unit-specific inefficiency estimates $\widetilde{E}(u|\varepsilon)$ ranged from 381.05 to 33074.41 with the mean and median being equal to 3883.34 and 3366.836 correspondingly. The 25^{th} percentile is 2396.84, and the 75^{th} percentile is 4538.75. The mode estimates ranged from 0 to 33074.41 with a mean of 3301.01 and median of 2947.18. The 25^{th} percentile is 1379.47, and 75^{th} percentile is 4403.69. Nonparametric density for unit-specific inefficiencies is presented on Figure 12.

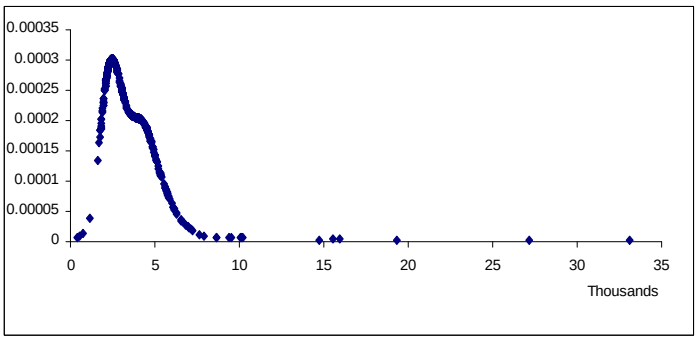

Figure 12. Nonparametric density for unit-specific inefficiency

The technical efficiency index (TEI) for the whole sample ranged from 0.0005 to 0.9305. The two highest efficiency scores belong to large scale combined cycle units.

The lowest scores are for small combustion turbines fueled by gas and a rather small combined cycle plant. The mean efficiency score is 0.14, the median is equal to 0.07. The 25^{th} and 75^{th} percentiles are 0.0143 and 0.20 correspondingly. Nonparametric density estimates for the scores are depicted on Figure 13.

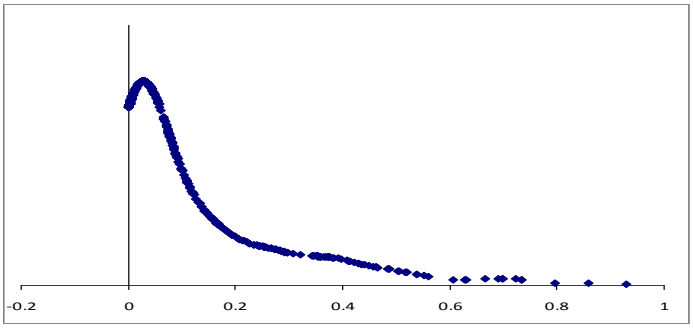

Figure 13. Nonparametric density estimates for technical efficiency

There are four observations with the highest TEI index, corresponding to the most efficient production units. For this group, the estimates for unit-specific inefficiency range from 381.05 to 1,611.51 and are the lowest across the sample. Mode estimates equal 0 for all four points. Technical efficiency indexes range from 0.73 to 0.93. Four top observations correspond to the combined cycle generation.

There are twelve observations with the most negative $\widetilde{\varepsilon_i}$ ranging from -40,725 to -10,686. They correspond to the highest unit-specific inefficiency estimates ranging from 8,679.85 to 33,074.41. With the exception of a small combined cycle plant, the rest of the observations in this group correspond to coal-fueled steam turbines. The mode estimates for these units also ranged from 8679 to 33,074 coming very close to the estimates of expected unit-specific inefficiency.

There are nine observations for which estimates of $\widetilde{\varepsilon_i}$ are above zero and the mode of the conditional distribution of the inefficiency term equals zero. Estimates of the unit-specific inefficiency scores range from 381.05 to 1,788.59.

The results of this analysis were grouped by fuel type. Figure 14 depicts the relationship between technical efficiency index estimates and type of generating unit. As before, this variable is broken down into the following categories: 0-combined cycle, 1 - gas-fueled combustion turbine , 2 - coal-fueled steam turbine, 3 - gas-fired steam generator and 4 - steam turbines using heavy fuel oil. Figure 15 shows unit-specific

inefficiency by type of generating unit.

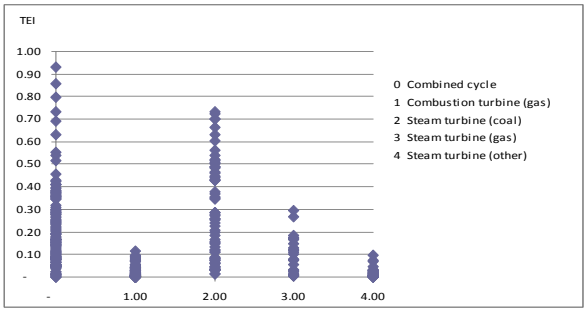

Figure 14. Scatter plot of TEI against the cycle types

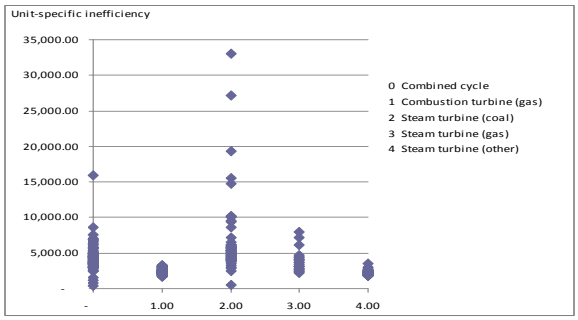

Figure 15. Scatter plot of unit-specific inefficiency

The output of the combined cycle generators ranges from 2205 to over 8 million MWh. Specific inefficiency scores for this group of generating units range from 381.05 to 15,921.51 with mean 4,255.75 and median of 4,126.56. The 25^{th} percentile is 3,510.00 and 75^{th} percentile equals 4,712.85. Mode estimates for the conditional distribution of inefficiency given composite error term range from 0 to 15,921.51 with a mean of 3,979.21 and median of 3,922.60. The 25^{th} and 75^{th} percentiles equal 3,142.62 and 4,600.00 correspondingly. The technical efficiency index for this type of generation ranged from 0.00074 to 0.93 with a mean of 0.21. The median TEI for combined cycle is 0.16, with the 25^{th} and 75^{th} percentile being equal to 0.08 and 0.29 respectively.

Estimates of $\widetilde{E}(u|\varepsilon)$ for gas-fueled combustion turbines ranged from 1,716.89 to 3,317.18 with the 25^{th} and 75^{th} percentile estimated at 2,151.91 and 2,615.39 cor-

respondingly. The mean and median equal 2,403.91 and 2,344.39 respectively. The technical efficiency index ranges from 0.0005 to 0.11. The mean is 0.02 and median is 0.013. The mode of the conditional distribution of u_i varies between 0 and 2,877.91. The 25^{th} and 75^{th} percentiles equal 876.38 and 1,782.02 correspondingly. The mean is 1,340.38 and median is 1,276.89.

Steam turbines across all fuel types (coal, gas and other) range in TEI from 0.001 to 0.73. The average TEI for steam turbines is 0.17, while the median is equal to 0.08. The 25^{th} and 75^{th} percentiles are 0.26 and 0.26 correspondingly. Unit-specific inefficiency scores for this category varied between 573.29 and 33,074.41 with a mean and median of 4,742.74 and 4,097.72 correspondingly. The bottom and top quartiles are 2,470.02 and 5,336.59. Mode estimates came to be close to the estimates of conditional distribution of u_i and ranged between 0 and 33,074.41. The mean is 4,245.80 and the median equals 3,887.91. The 25^{th} percentile is 1,518.52 and the 75^{th} percentile is 5,279.13.

Coal fueled steam turbines seem to have the highest unit-specific inefficiency scores out of the group. The mean value of technical inefficiency score is 6,366.25, the median is 5,181.09. Minimum and maximum values are 573.30 and 33,074.41 correspondingly. The 25^{th} percentile is 4,228.56 and the 75^{th} percentile is 5,814.81. The average technical efficiency index for this category is 0.27 and the median is 0.22. The estimates range from 0.013 to 0.92. Estimates of the mode for the conditional distribution of inefficiency given the composite error term are not very different from the expected individual technical inefficiency.

Steam turbines that fueled by gas have significantly lower individual inefficiency scores than the coal-fired turbines. The mean is 3,839.70 and median is 3,534.02. The estimates vary between 2,156.89 and 7,911.41 with the 25^{th} and 75^{th} percentiles equaling 2,899.89 and 4,240.57 respectively. Again, the estimates of the mode for the conditional distribution of u_i given ε_i are similar to the unit-specific technical inefficiency scores. The technical efficiency index varies between 0.005 and 0.30. The mean TEI for this category is 0.10,and the median is 0.09.

The last category of steam turbines, those that use other fuel, has the lowest estimates of unit-specific inefficiency scores across all steam-driven generators. The mean is 2,457.02, and the median is 2,228.68 with the estimates varying in the interval between 1,786.92 and 3,533.65. The technical efficiency index for this group ranges between 0.0006 and 0.0957 with mean being equal to 0.027 and the median estimated

at 0.0085. Mode estimates for the conditional distribution of u_i range between 0 and 3,174.32, the 25^{th} percentile equals 545.29 and the 75^{th} percentile equals 1,506.72. The mean and median are 1,461.75 and 1,040.96 respectively.

Overall, gas-fueled turbines, turbines fueled by "other" fuel and combustion turbines have the lowest unit-specific inefficiency scores among all the groups. They also had the lowest TEI indices across all groups. Out of the high TEI units, average for the coal fired turbines is close to the average TEI for the combined cycle turbines. But the average and median unit-specific inefficiency scores are lower for the combined cycle generators by 2,110.49 and 1,054.53 respectively.Confidence intervals for the unit-specific inefficiency estimates are depicted in Figure 16. Confidence intervals (CI) were constructed based on Greene (2006). CI is not interpreted by Greene (2006) as true confidence interval, but rather as a range that includes $100(1-\alpha)\%$ of the conditional distribution of u_i given ε_i, where α is a significance level. Also, one-way hypothesis testing for unit-specific efficiency scores is discussed in Bera and Sharma (1999). The 95% confidence limits are computed as

$$\begin{aligned} LB_i &= \mu_i^* + \sigma^* \Phi^{-1}[1-(1-\frac{\alpha}{2})\Phi(\frac{\mu_i^*}{\sigma^*})], \\ UB_i &= \mu_i^* + \sigma^* \Phi^{-1}[1-\frac{\alpha}{2}\Phi(\frac{\mu_i^*}{\sigma^*})] \end{aligned}$$

where LB_i and UB_i are lower and upper bound respectively, $\mu_i^* = -\varepsilon_i \lambda^2/(1+\lambda^2)$ and $\sigma^* = \sigma\lambda/(1+\lambda^2)$.

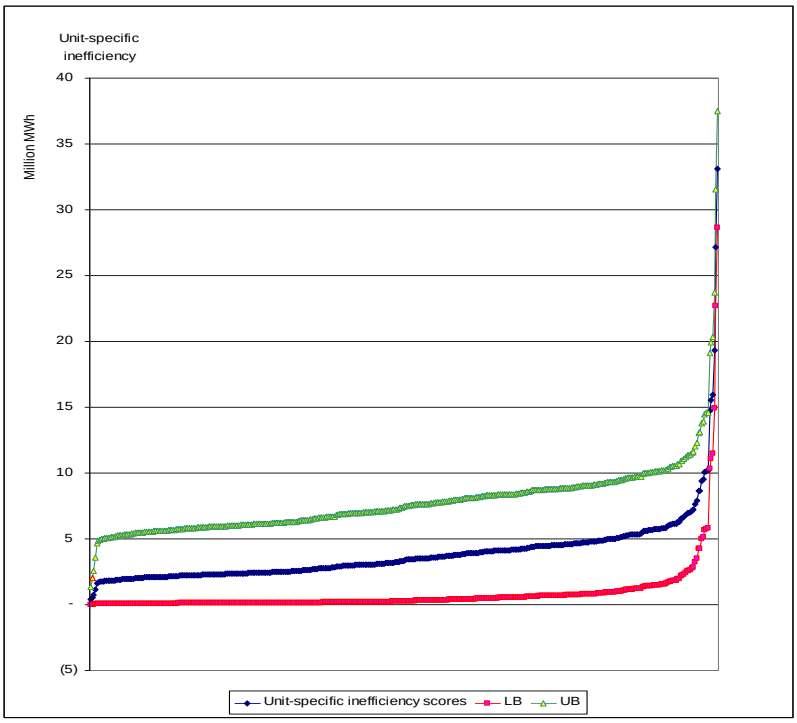

Figure 16. 95% Confidence intervals for unit-specific inefficiency scores

This comparison within the limitations of available data and applied estimation procedure suggests that coal steam turbines have higher average unit-specific inefficiency scores than the combustion cycle, while their TEI estimates are close. The remaining three turbines have smaller unit-specific inefficiency scores, but their technical efficiency indexes are low as well. It is worth noting that discussion of the power generation efficiency in the present study should only be viewed in the context of limitations of the imposed assumptions, utilized estimation procedure and available data.

3.5 Conclusion

This essay attempted to analyze efficiency of generating units by establishing a benchmark ranking, which would be used in the future to study changes in productivity and efficiency under different emission mitigation scenarios. We analyzed a model that combines a semiparametric stochastic frontier estimation with additive restriction on the production function. Categorical variable accounting for the variability in cycle and fuel type was treated in the context of kernel smoothing, which is

different from traditionally used frequency approach. Discussion of empirical results concluded the analysis.

Fairly flexible estimation procedure utilized here generated results consistent with engineering findings that natural gas combined cycle generation is more efficient than combustion or steam-turbines in the context of economic productivity and efficiency measurement. Efficiency scores here have different interpretation from the engineering definition of efficiency. The latter one is based on the heat rate of a generator and indicates how efficiently a generator converts energy from burning fuel into the electricity. Efficiency scores studied in the context of frontier estimation have different interpretation. They represent overall proximity to the production frontier when all other inputs and costs are taken into the account. Therefore it contains information on how well a production units transform all inputs into the final output. In addition, stochastic frontier incorporates the external shocks that are beyond control of the generating unit. Firm-specific inefficiency scores derived in this context allow production unit to be inefficient relative to its maximum possible level of output given by its stochastic frontier, which already accounts for the effect of external random factors outside the control that are not observed directly.

4 General conclusion

This research focused on applying semiparametric methods to analyze both the demand for natural gas in the residential housing sector and efficiency of electricity generation. The first essay investigated the relationship between natural gas demand and characteristics of the dwelling, demographic characteristics of occupants and behavioral variables. The existing modeling literature, whether it relies on parametric specifications or engineering simulation, does not accommodate inclusion of a behavioral component. This essay attempts to bridge that gap and investigate the applicability of additive nonparametric regression to this task. The results of this analysis can be used for three primary purposes. The first one is an in-sample prediction for approximating energy demand of a residential building whose characteristics are described by the regressors in this analysis, but a certain combination of their particular values does not exist in the real world. The second potential application is for benefit-cost analysis of residential upgrades and retrofits under a fixed budget, since the results of this study contain information on how natural gas consumption might change once a particular characteristic or attribute is altered. The third purpose is to establish a relationship between natural gas consumption and changes in behavior of occupants. Although information on behavioral variables is generally limited, results of the analysis identify what information would be helpful to further research.

The second essay studies the efficiency of power generation for five types of generating units: combined cycle, combustion turbine and three kinds of steam turbine. This study is different from previously published research as it contains an application of a semiparametric stochastic frontier model with the introduction of an additivity constraint on the production function of electricity generation. The semiparametric model of Fan et al. (1996) is used as a foundation for this analysis, but this model takes it a step further by incorporating an additivity restriction on the production function. This is handled nonparametrically relying on the smooth backfitting framework of Mammen et al. (1999). In addition, the methodology includes kernel smoothing of the categorical variables inside of the smooth backfitting estimation. This essay incorporates some of the advances in non-parametric estimation and results in a fully operational estimation procedure with a straight-forward algorithm, which is capable of nonparametrically handling categorical variables to model the production function and analyze efficiency in the context of stochastic frontiers.

5 Bibliography

1. Aigner, D.J., C.A.K. Lovell and P.J. Schmidt, 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6, 21-37.

2. Aitchison, J. & Aitken, C.G.G., 1976. Multivariate binary discrimination by the kernel method. Biometrika 63, 413-420.

3. Baker, P., Blundell, R. W. and Micklewright, J., 1989. Modelling household energy expenditures using micro-data. Economic Journal 99, 720-738.

4. Baltagi, B.H. and J. M. Griffin, 1988. A General index of technical change. The Journal of Political Economy 96, 20-41.

5. Battese, G. E., Corra, G. S., 1977. Estimation of a production frontier model with application to the pastoral zone of Eastern Australia. Australian Journal of Agricultural Economics 21, 169-179.

6. Bera, A.K., Sharma S.C., 1999. Estimating production uncertainty in stochastic frontier production function models. Journal of Productivity Analysis 12, 187–210

7. Buja, A., Hastie, T. and Tibshirani, R., 1989. Linear smoothers and additive models. Annals of Statistics 17, 453–510.

8. Callan, S. J., 1991. The sensitivity of productivity growth measures to alternative structural and behavioral assumptions: an application to electric utilities, 1951–1984. Journal of Business and Economic Statistics 9, 207–213.

9. Caravan Opinion Research Corporation, 2007. Study #716287.

10. Crawley, Drury B, Linda K Lawrie, Curtis O Pedersen, Frederick C Winkelmann, Michael J Witte, Richard K Strand, Richard J Liesen, Walter F Buhl, Yu Joe Huang, Robert H Henninger, Jason Glazer, Daniel E Fisher, Don B Shirey III, Brent T Griffith, Peter G Ellis, Lixing Gu. 2004. EnergyPlus: New, Capable, and Linked. Journal of Architectural and Planning Research 21, 4 (Winter 2004).

11. Deprins, D., L. Simar and H. Tulkens, 1984. Measuring labor inefficiency in post offices, in: M. Marchand, P. Pestiau and H. Tulkens, (Eds.), The performance of public enterprises: concepts and measurements. North Holland, Amsterdam.

12. DOE-2, 1993. BDL Summary Version 2.1E. LBL, 34946, Lawrence Berkeley National Laboratory, Berkeley, CA

13. Dorfman, J. H. and Atkinson, S.E., 2005. Bayesian measurement of productivity and efficiency in the presence of undesirable outputs: crediting electric utilities for reducing air pollution. Journal of Econometrics 126, 445-468.

14. Fan, Y., Q. Li, and A. Weersink, 1996. Semiparametric estimation of stochastic production frontier models. Journal of Business and Economic Statistics 14, 460-468.

15. Farrell, M.J., 1957. The measurement of production efficiency. Journal of the Royal Statistical Society. Series A 120, 253-81.

16. García-Cerruti, L., 2000. Estimating elasticities of residential energy demand from panel county data using dynamic random variables models with heteroskedastic and correlated error terms. Resource and Energy Economics 22, 355-366.

17. Gauss 9.0 User Guide, 2009. Aptech Systems, Inc.

18. Greene, W.H., 1990. Maximum likelihood estimation of econometric frontier, Journal of Econometrics 13, 27–56.

19. Greene, W.H., 2006. The econometric approach to efficiency analysis in K. Lovell and S. Schmidt, eds. "The measurement of efficiency", H. Fried. Oxford University Press

20. Griffin, J. E., and Steel, M. F. J.,2004. Semiparametric Bayesian inference for stochastic frontier models. Journal of Econometrics 123, 121-152.

21. Hall, P., J.S. Racine, and Q. Li, 2004. Cross-Validation and the Estimation of Conditional Probability Densities. Journal of the American Statistical Association 99, 1015-1026.

22. Halvorsen, B. and B. Larsen, 2001. The flexibility of household electricity demand over time." Resource and Energy Economics 23, 1-18.

23. Härdle, W., Hall, P. and Marron, J. S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). Journal of the American Statistical Association 83, 86–101.

24. Henderson, D.J. and L. Simar, 2005. A fully nonparametric stochastic frontier model for panel data. Discussion Paper 0417, Institute de Statistique, Universite Catholique de Louvain.

25. Hiebert, L. D., 2002. The determinants of the cost efficiency of electric generating plants: a stochastic frontier production approach. Southern Economic Journal 68, 935-946.

26. Holtedahl, P. and F. Joutz, 2004. Residential electricity demand in Taiwan. Energy Economics 26, 201-224.

27. Jondrow, J., K. Lovell, C. A., Materov, Ivan S. and Schmidt, P.,1982. On the estimation of technical inefficiency in the stochastic frontier production function model. Journal of Econometrics 19, 233-238.

28. Kamerschen, D. and D. Porter, 2004. The demand for residential, industrial and total electricity,1973-1998. Energy Economics 26, 87-100.

29. Kleit, A. and Terrell D., 2001. Measuring potential efficiency gains from deregulation of electricity generation: a Bayesian approach. The Review of Economics and Statistics 83, 523-530.

30. Kumbhakar, S.C., B.U. Park, L. Simar and E.G. Tsionas, 2007. Nonparametric stochastic frontiers: A local maximum likelihood approach. Journal of Econometrics 137, 1-27.

31. Kumbhakar, S.C., Tsionas E. G. and Sipiläinen, T., 2009. Joint estimation of technology choice and technical efficiency: an application to organic and conventional dairy farming. Journal of Productivity Analysis 31, 151-161.

32. Labandeira, X., J. M. Labeaga, and M. Rodriguez, 2006. A residential energy demand system for Spain. The Energy Journal 27, 87-111.

33. Labandeira, X., Labeaga, J.M. and M. Rodríguez, 2004. Microsimulating the effects of household energy price changes in Spain. Fondazione Eni Enrico Mattei, Working Paper # 161.

34. Larsen, B. and R. Nesbakken, 2004. Household electricity end-use consumption: results from econometric and engineering models. Energy Economics 26, 179-200.

35. Li, Q. & Racine, J., 2003. Nonparametric estimation of distributions with categorical and continuous data. Journal of Multivariate Analysis 86, 266-292.

36. Li, Q. and J. S. Racine, 2004. Cross-validated local linear nonparametric regression. Statistica Sinica 14, 485-512.

37. Li, Q. and J. S. Racine, 2007. Nonparametric econometrics: theory and practice. Princeton University Press, 768.

38. Linton, O. and J.P. Nielsen, 1995. A kernel method of estimating structured nonparametric regression based on marginal integration, Biometrika 82, 93-100.

39. Lutzenhiser, L., 1993. Social and behavioral aspects of energy use. Annual Review of Energy Economics 18, 247-89.

40. Madlener, R., 1996. Econometric analysis of residential energy demand: a survey. Journal of Energy Literature 2, 3-32.

41. Mammen, E. and Park, B. U., 2005. Bandwidth selection for smooth backfitting in additive models. Annals of Statistics 33, 1260–1294.

42. Mammen, E., Linton, O. and Nielsen, J. P., 1999. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Annals of Statistics 27, 1443–1490.

43. Martins-Filho, C and Yao F., 2009. Nonparametric stochastic frontier estimation via profile likelihood. Working paper, Department of Economics, University of Colorado, Boulder.

44. Martins-Filho, C., 2006. Applied microeconomics: course notes. Department of Economics, Oregon State University.

45. McDonnell, J. T., 1991. Wholesale power substitution for fossil and nuclear fuels by electric utilities: a cross-sectional analysis. Master thesis. Golden, Mineral Economics Department, Colorado School of Mines.

46. Meeusen, W., and J. Van den Broeck. 1977. Efficiency estimation from a Cobb-Douglas production function with composed error. International Economic Review 18, 435-444.

47. Narayan, P. and R. Smyth, 2005. The residential demand for electricity in Australia: an application of the bounds testing approach to cointegration. Energy Policy 33, 467-474.

48. Nelson, R. A. ,1984. Regulation, capital vintage, and technical change in the electric utility industry. Review of Economics and Statistics 66, 59–69.

49. Nesbakken, R., 2001. Energy consumption for space heating: discrete-continuous approach. Scandinavian Journal of Economics 103, 165-184.

50. Newey, W.,1994. Kernel estimation of partial means. Econometric Theory 10, 233-253.

51. Nielsen J.P., Sperlich S., 2005. Smooth backfitting in practice. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 43-61.

52. Opsomer, J. D. and Ruppert, D., 1997. Fitting a bivariate additive model by local polynomial regression. Annals of Statistics 25, 186–211.

53. Opsomer, J. D., 2000. Asymptotic properties of backfitting estimators. Journal Multivariate Analysis 73, 166–179.

54. Pasurka, C., 2003. Changes in emissions from U.S. manufacturing: a joint production perspective. Social Science Research Network.

55. PROMOD IV, 2008. Ventyx.

56. Racine, J. S. and Q. Li, 2004. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics 119, 99-130.

57. Racine, J.S., Q. Li., and X. Zhu, 2004. Kernel estimation of multivariate conditional distributions. Annals of Economics and Finance 5, 211-235.

58. Schmalensee, R. and Stoker, T. M.,1999. Household gasoline demand in the United States. Econometrica 67, 645-662.

59. Seiford L. M. and Thrall R.M., 1990. Recent developments in DEA : The mathematical programming approach to frontier analysis. Journal of Econometrics 46, 7-38.

60. U.S. Department of Energy, 2005. ENERGYPLUS, Input Output Reference.

61. U.S. Department of Energy, Energy Information Administration (DOE/EIA). 2009. Annual energy outlook 2009 with projections to 2030. DOE/EIA-0383(2009).

62. Wang, M.C., and J. Ryzin, 1981. A class of smooth estimators for discrete distributions. Biometrika 68, 301-309.

63. Yatchew, A. and J. No, 2001. Household gasoline demand in Canada. Econometrica 69, 1697-1709.

Appendix

Appendix A. Charts for directional regression results

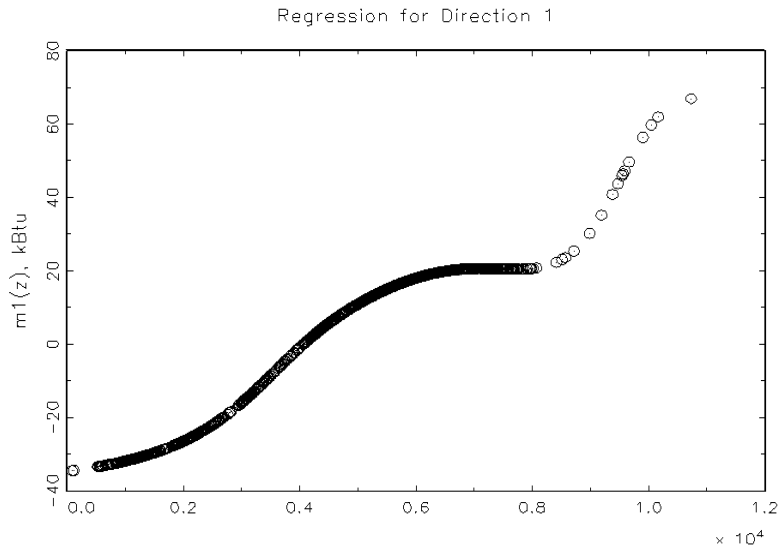

Figure A.1 Heating degree days: base=65, 01 to 12-2005

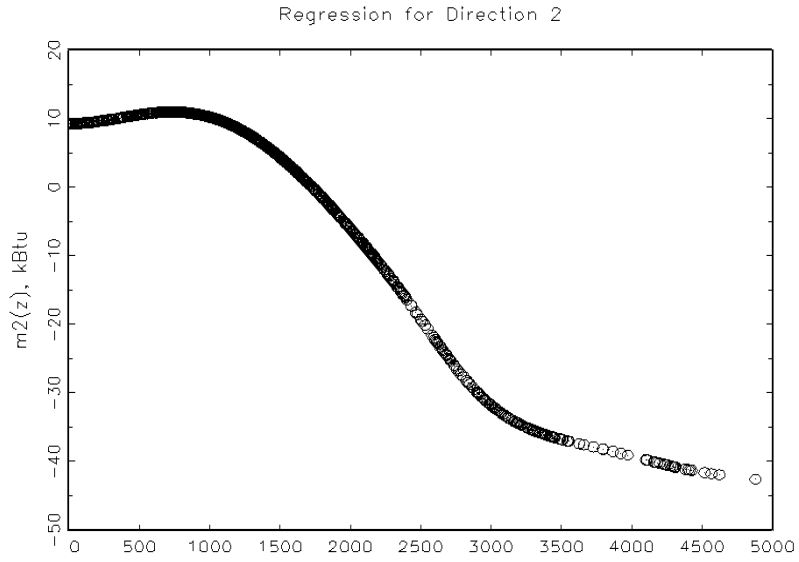

Figure A.2 Cooling degree days: base=65, 01 to 12-2005

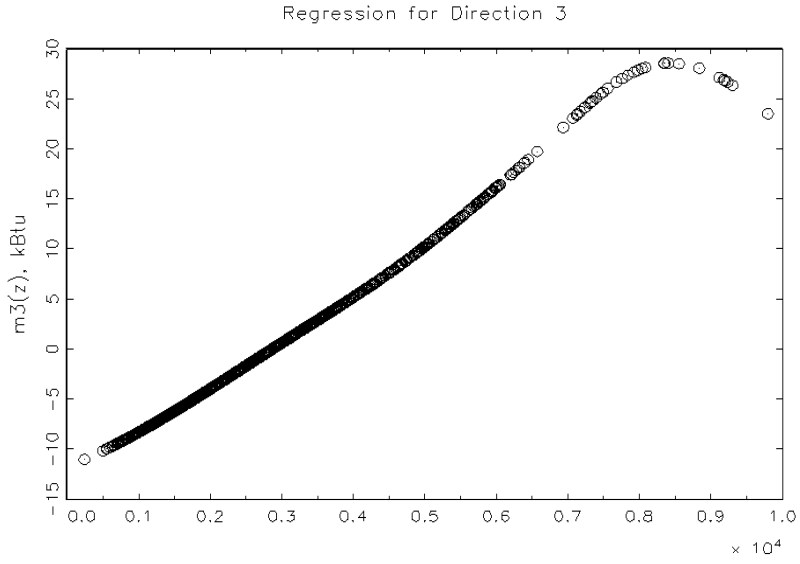

Figure A.3 Total house area

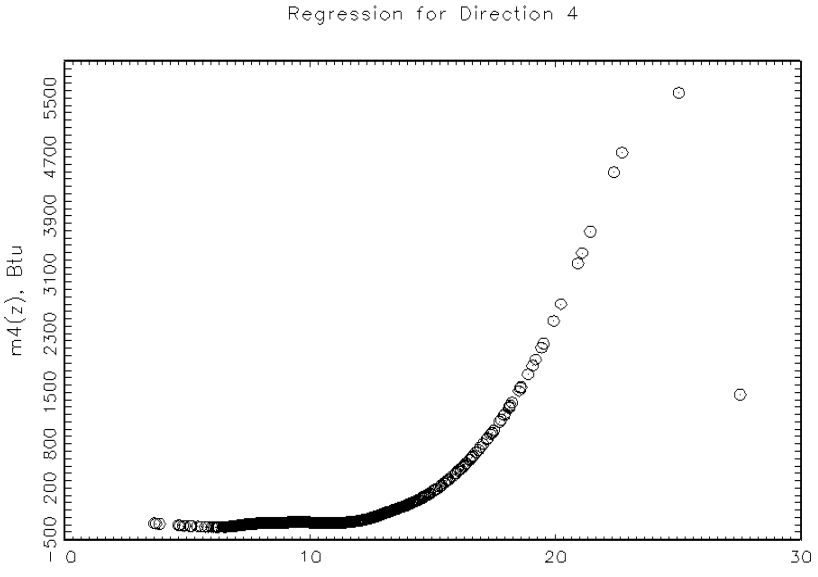

Figure A.4 Price of electricity, cents/KWh

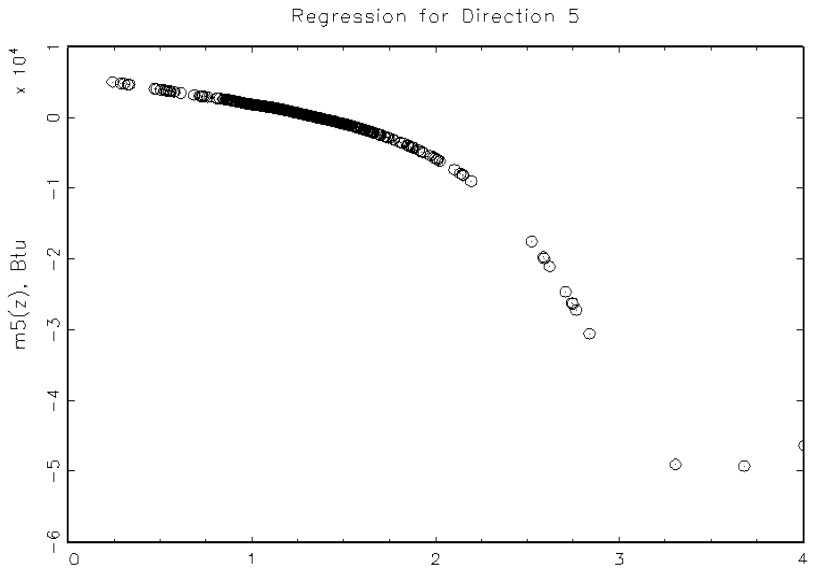

Figure A.5 Price of natural gas, cents*10/Btu

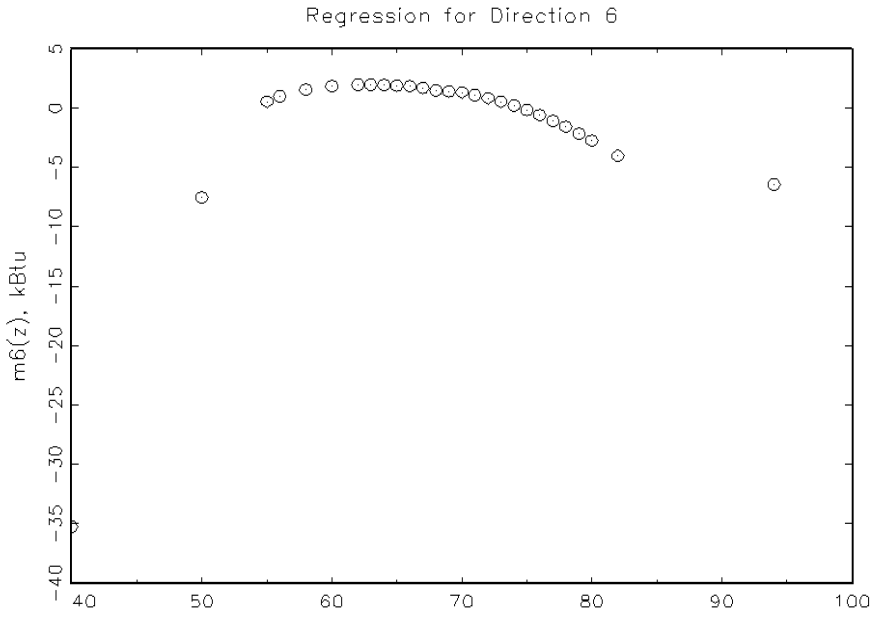

Figure A.6 Setting during the winter day when someone is home

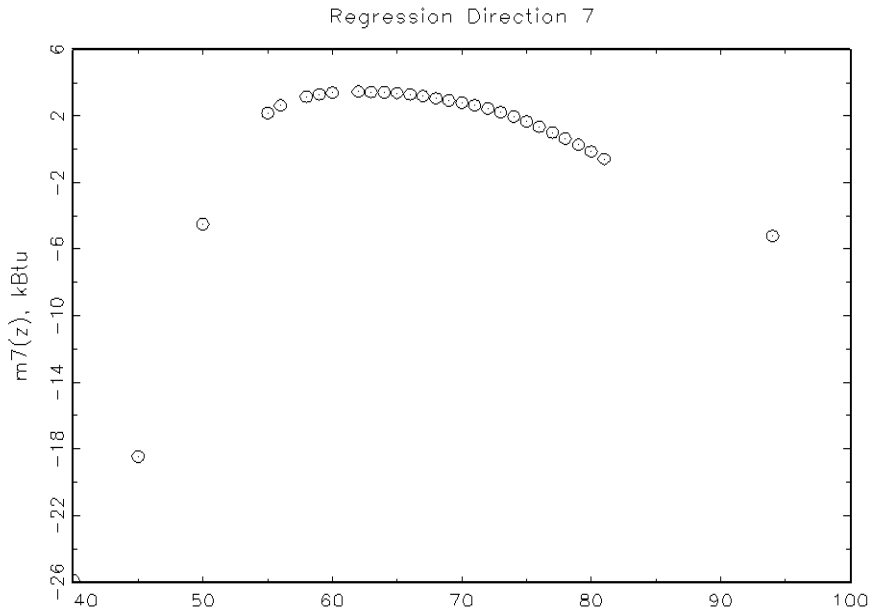

Figure A.7 Setting during the winter day when no one is home

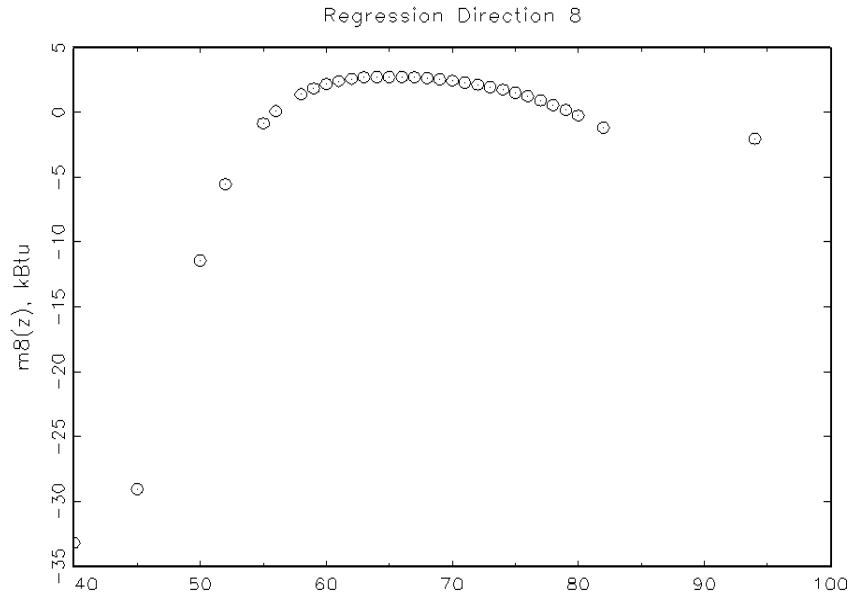

Figure A.8 Setting during sleeping hours in winter

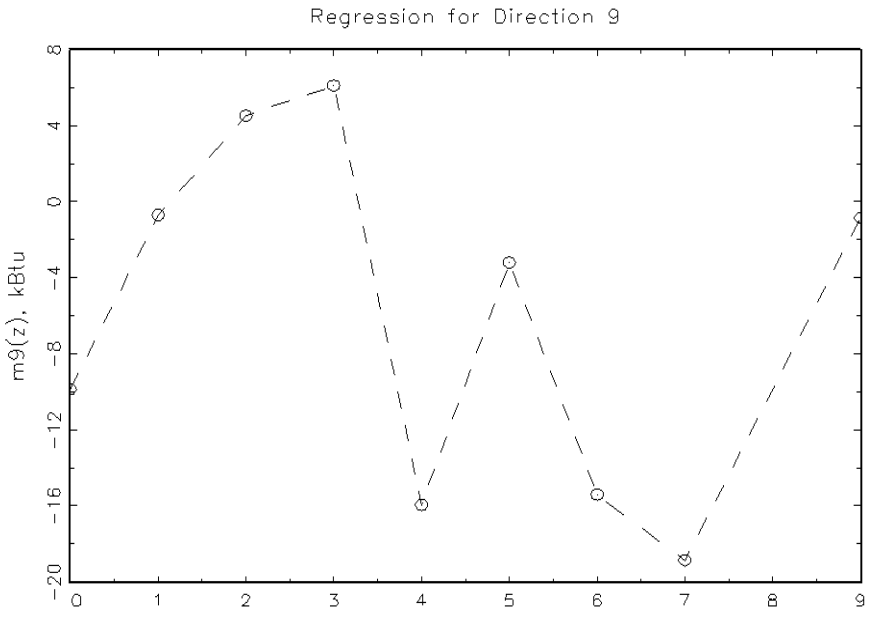

Figure A.9 Exterior wall construction material

0 Indescribable
1 Brick
2 Wood
3 Siding (Aluminum, vinyl, or steel)
4 Stucco
5 Composition (Shingle)
6 Stone
7 Concrete or concrete block
8 Glass
9 Other

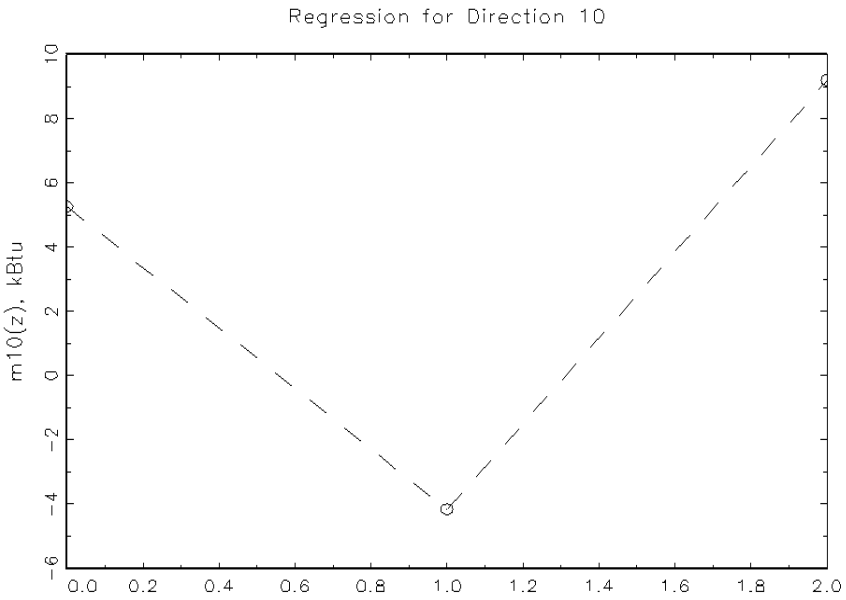

Figure A.10 Is the garage heated

0 No garage

1 Not heated

2 Yes

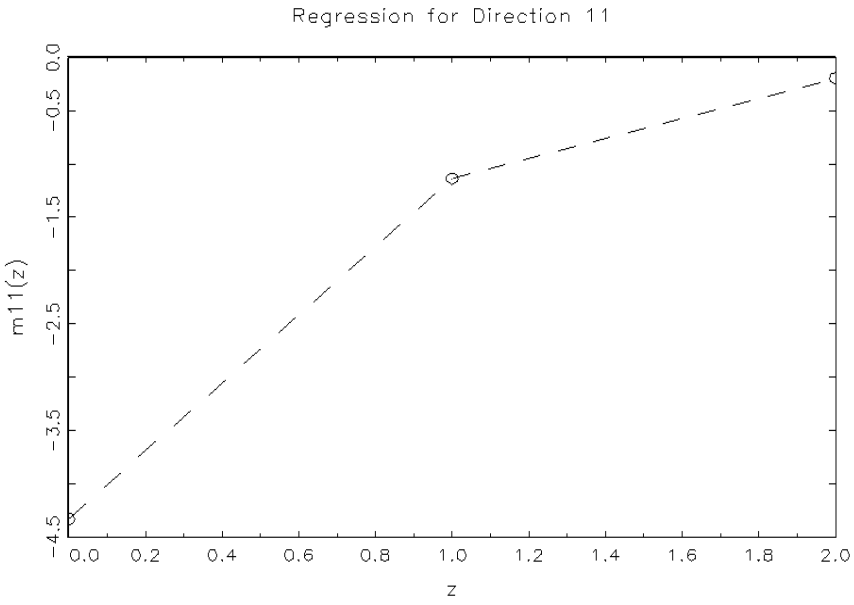

Figure A.11 Dwelling owned or rented

0 Own

1 Rent

2 Occupied w/out payment

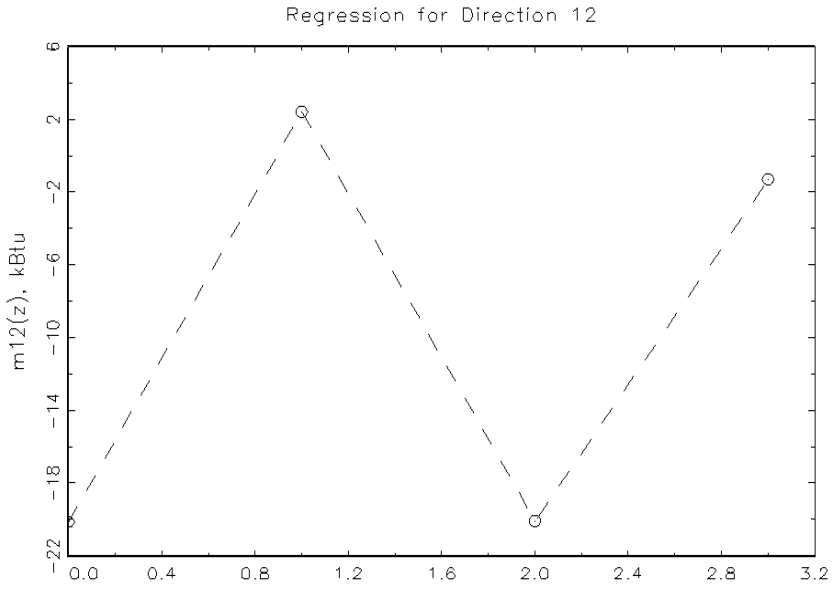

Figure A.12 Fuel used by cooking burners

0 Some other fuel

1 Natural gas from underground pipes,

2 Propane (bottled gas), or

3 Electricity

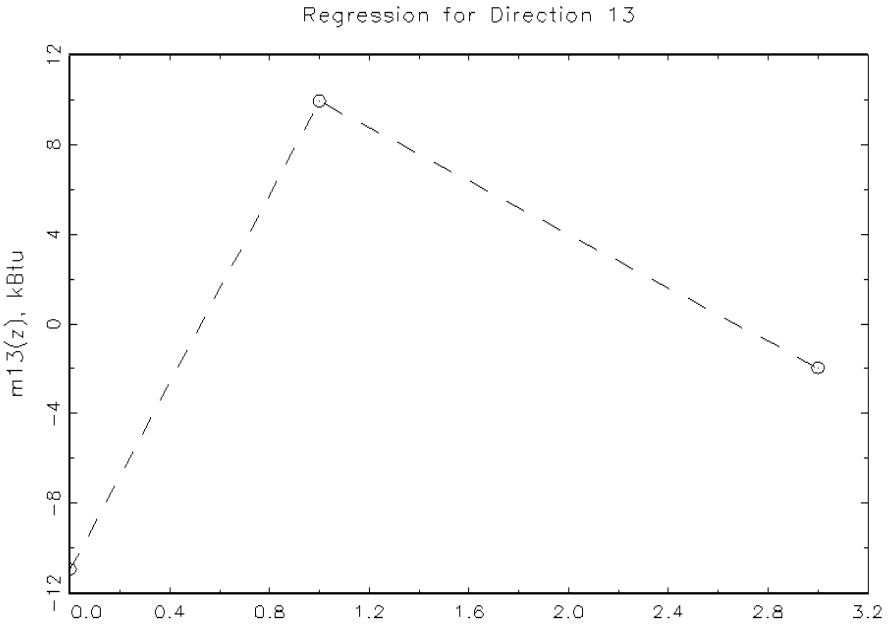

Figure A.13 Fuel used by clothes dryer

0 No dryer

1 Natural gas from underground pipes,

2 Propane (bottled gas), or

3 Electricity,

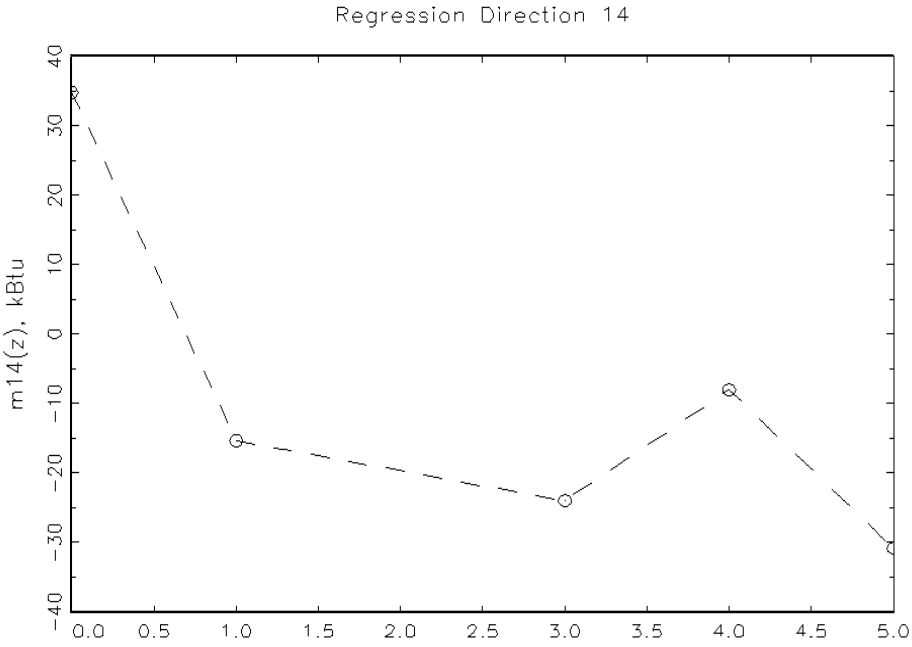

Figure A.14 Combined all secondary heating equipment

0 No secondary heating equipment

1 Central warm-air furnace with ducts to individual rooms other than a heat pump

2 Steam/hot water system with radiators/convectors in each room or pipes in the floor or walls

3 Built-in floor/wall pipeless furnace

4 Built-in room heater burning gas, oil, or kerosene

5 Cooking stove used for heating and cooking

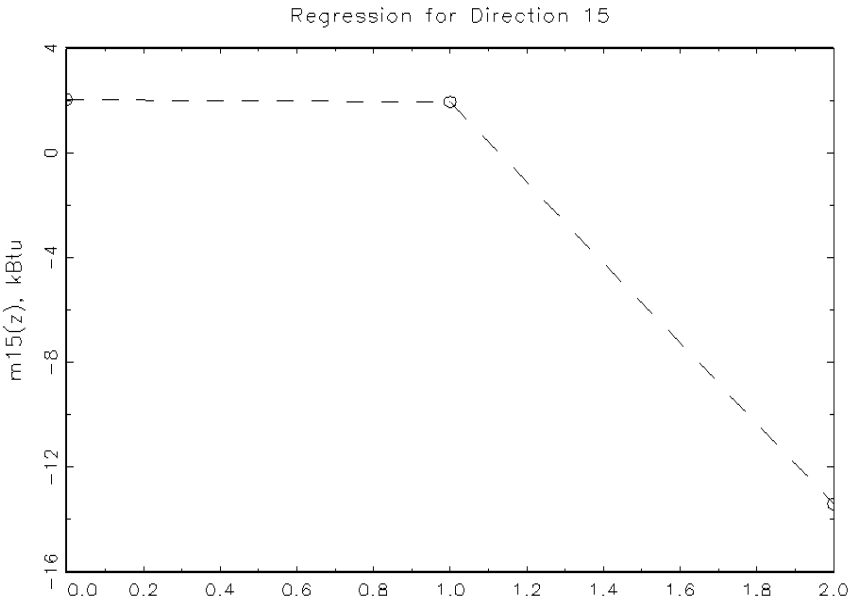

Figure A.15 Is the thermostat programmable

0 No

1 Yes

2 No thermostat

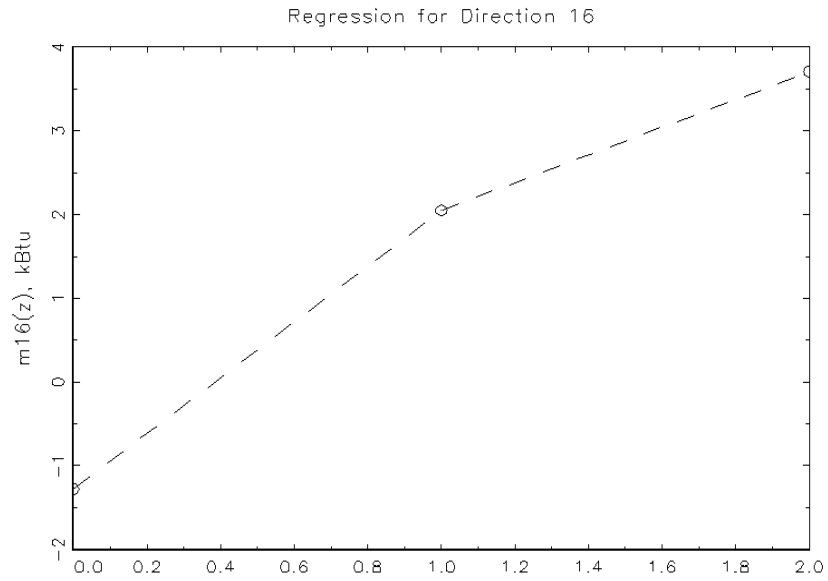

Figure A.16 Programmable thermostat lowers heat at night

0 No

1 Yes

2 No thermostat or not programmable

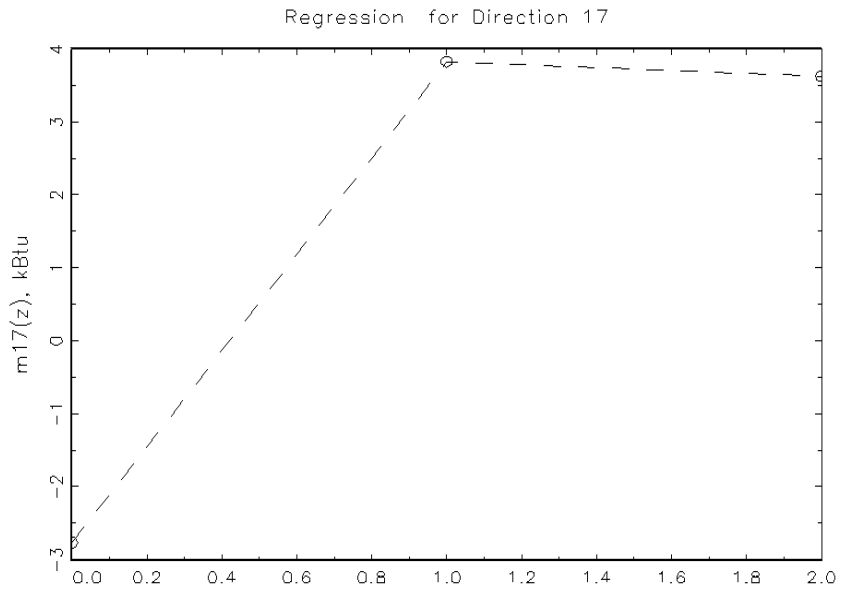

Figure A.17 Programmable therm lowers heat during the day

0 No

1 Yes

2 No thermostat or not programmable

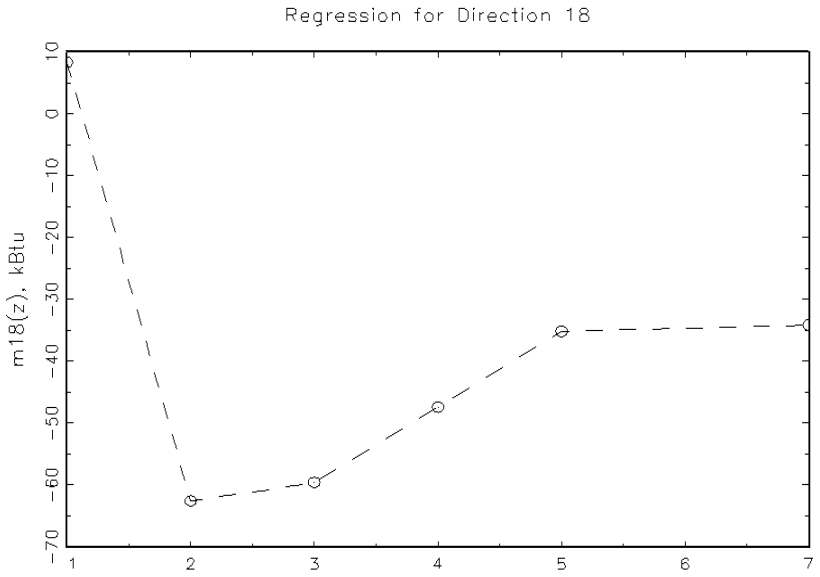

Figure A.18 Main fuel used for heating home

0 Propane (bottled gas)

1 Natural gas from underground pipes

2 Fuel oil

3 Kerosene

4 Electricity

5 Wood

7 Solar

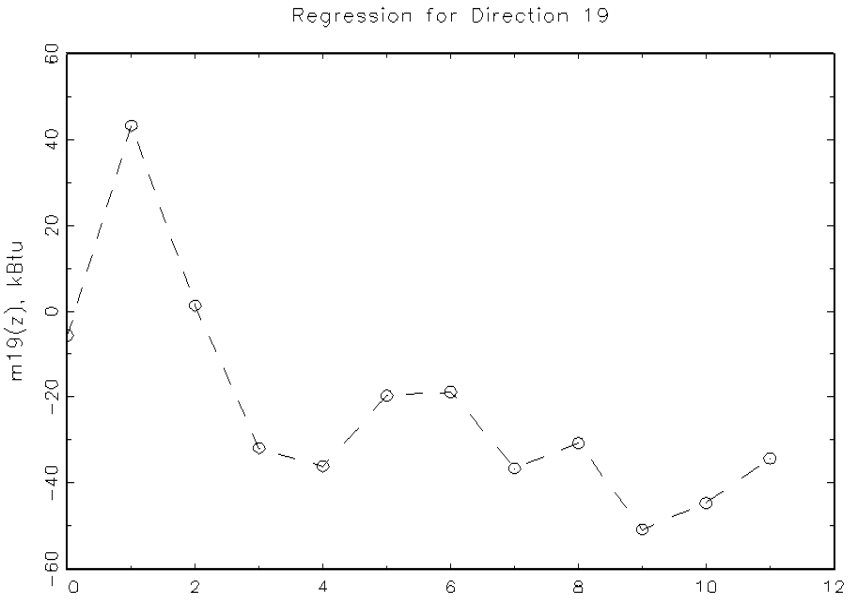

Figure A.19 Type of heating equipment providing the heat

0 No heating equipment used

1 Steam/Hot water system with radiators/convectors in each room or pipes in the floor or walls

2 Heat pump

3 Central warm-air furnace with ducts to individual rooms other than a heat pump

4 Built-in electric units in each room installed in walls, ceiling, baseboard, or floor

5 Built-in floor/wall pipeless furnace

6 Built-in room heater burning gas, oil, or kerosene

7 Heating stove burning wood, coal, or coke

8 Fireplace

9 Portable electric heaters

10 Portable kerosene heaters

11 Cooking stove that is used for heating and cooking

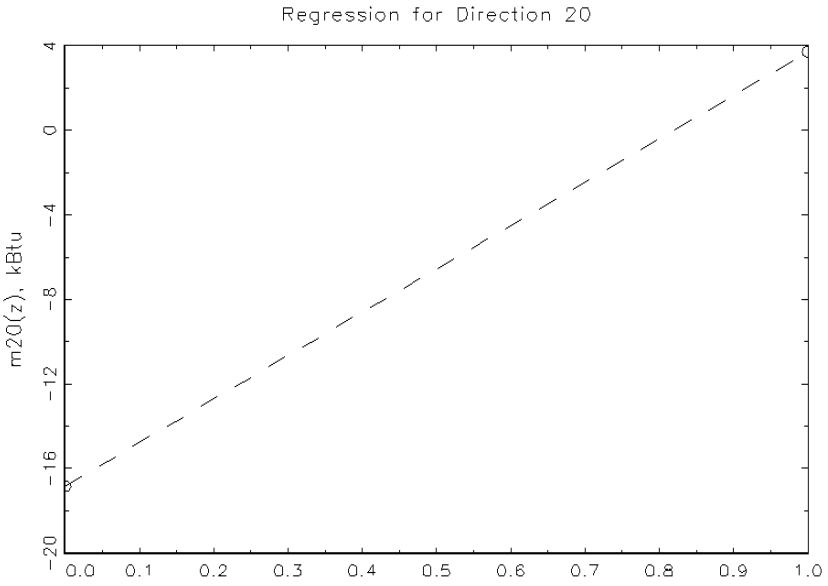

Figure A.20 Natural gas used for water heating

0 No

1 Yes

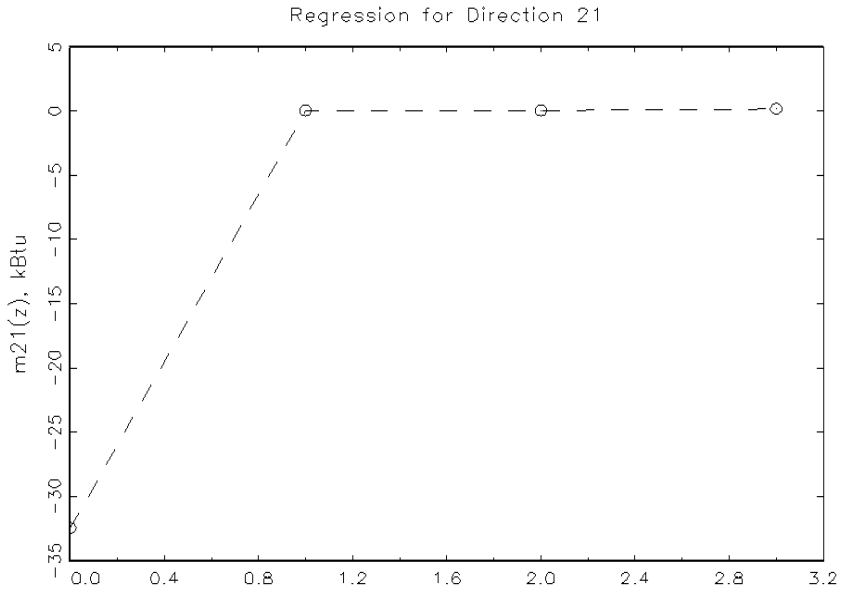

Figure A.21 How natural gas is paid

0 HH pays all

1 All in Rent/Fee

2 Some paid, some included in rent

3 Other

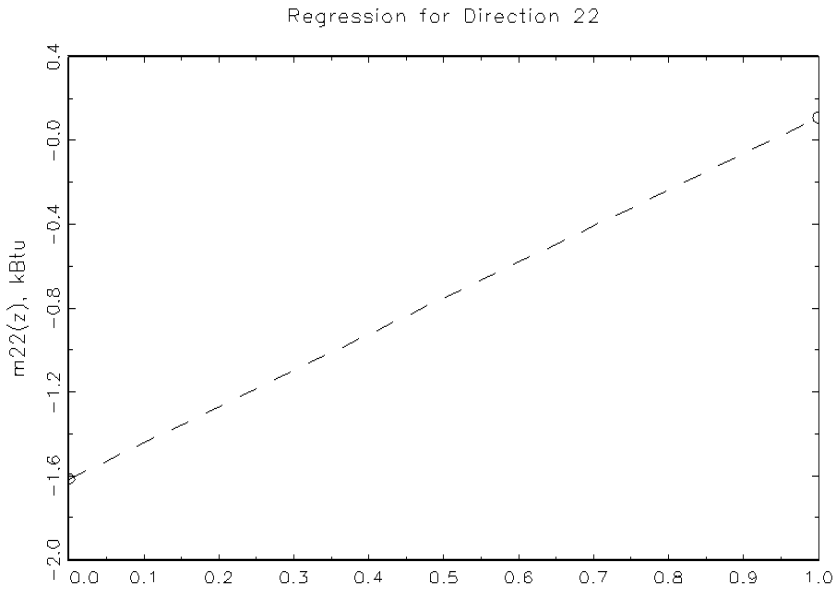

Figure A.22 Is someone at home all day on a typical weekday

0 No

1 Yes

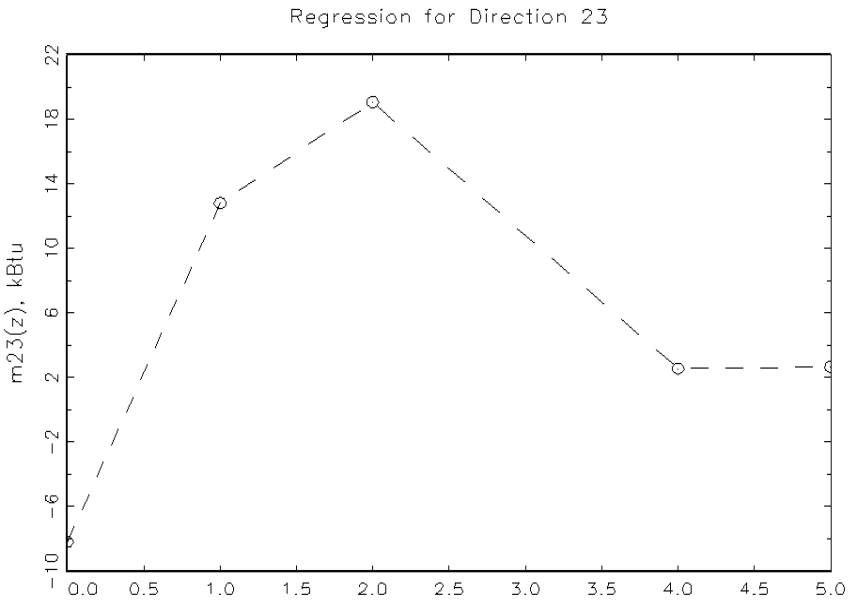

Figure A.23 Reported stories in housing unit

0	One story
1	Two stories
2	Three stories
3	Four or more
4	Split level
5	Other

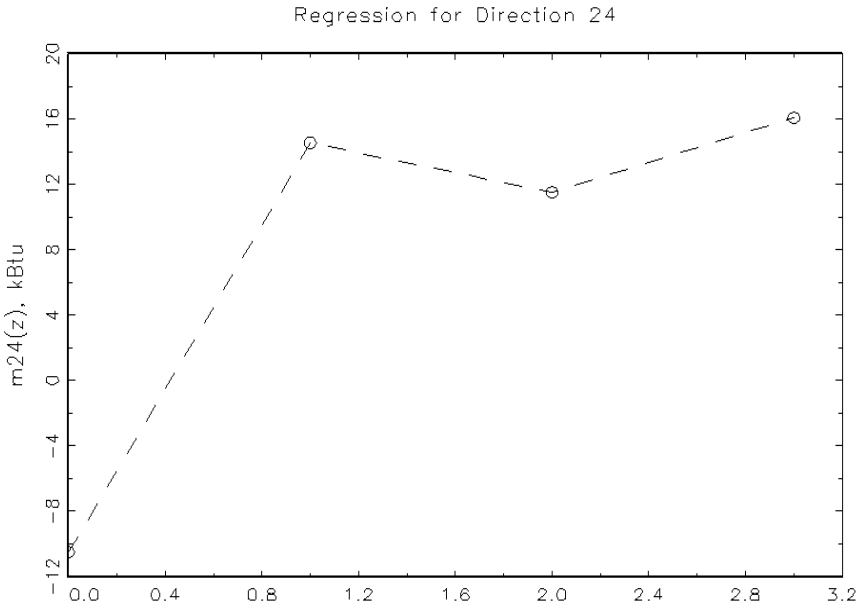

Figure A.24 Basement/crawl space heated

0 no basement

1 none

2 part

3 all

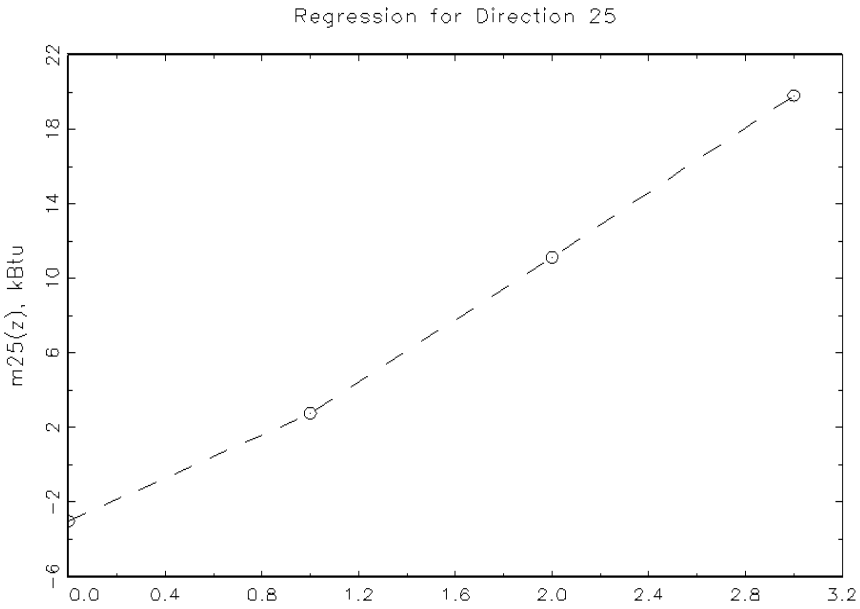

Figure A.25 How much of the attic is warm

0 no attic

1 none

2 part

3 all

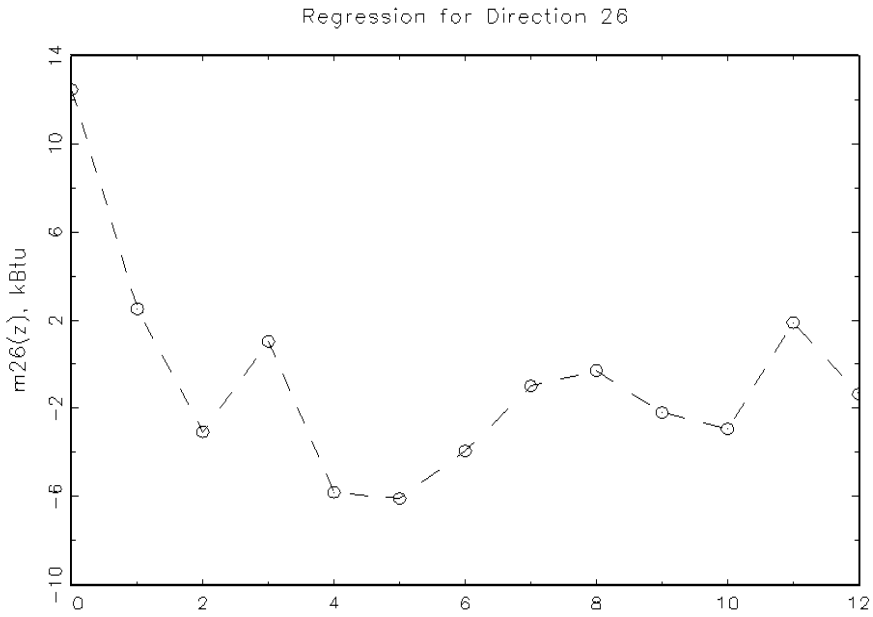

Figure A.26 Year home built

0	BEFORE 1940
1	1940-49
2	1950-59
3	1960-69
4	1970-79
5	1980-84
6	1985-89
7	1990-94
8	1995-99
9	2000-02
10	2003
11	2004
12	2005

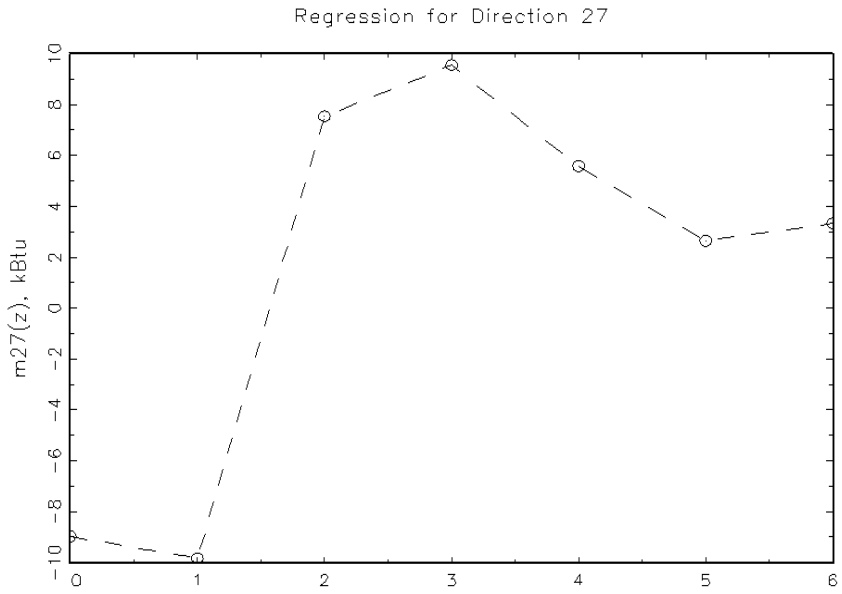

Figure A.27 How many thermostats overall

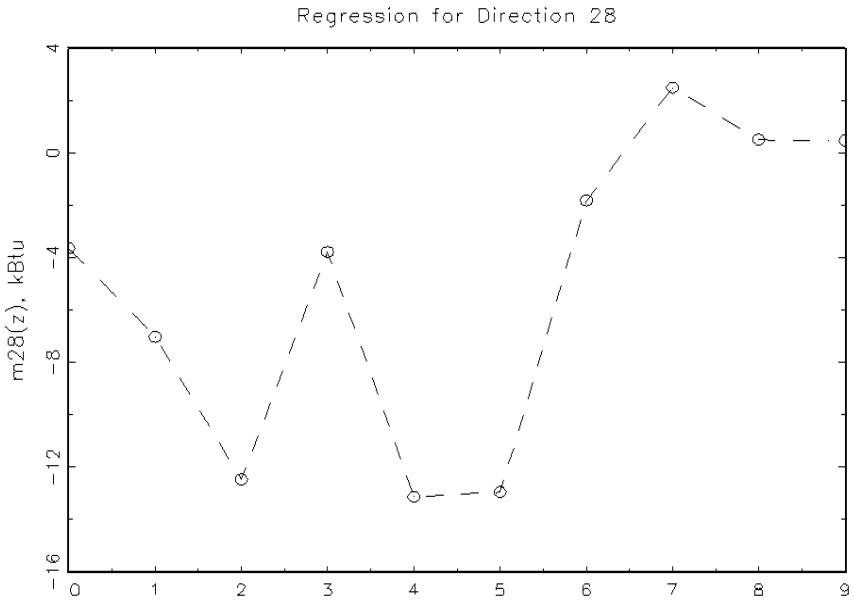

Figure A.28 Number of rooms not heated last winter

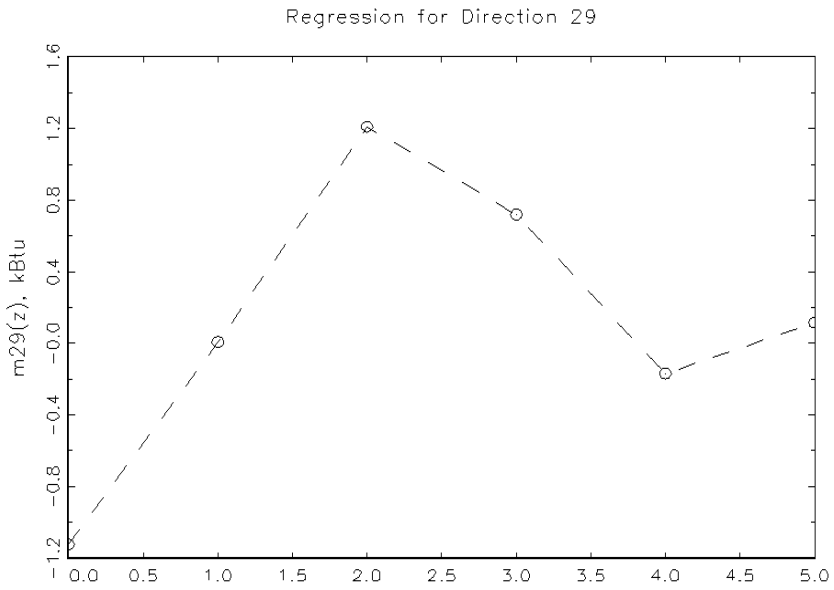

Figure A.29 Type of window glass

0 Single-pane glass

1 Double-pane glass

2 Double-pane glass with Low-E coating

3 Triple-pane glass

4 and 5 Triple-pane glass with Low-E coatings

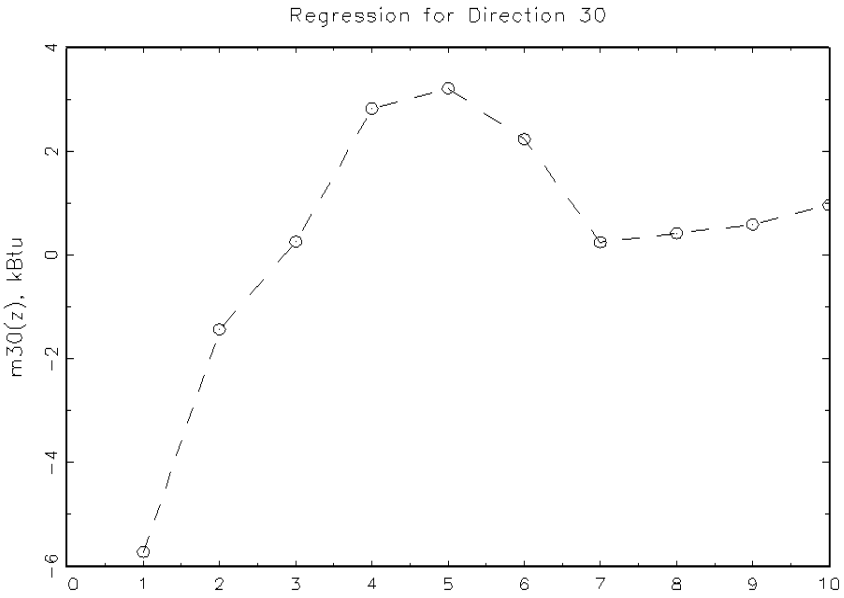

Figure A.30 Number of occupants (0=none, up to 10)

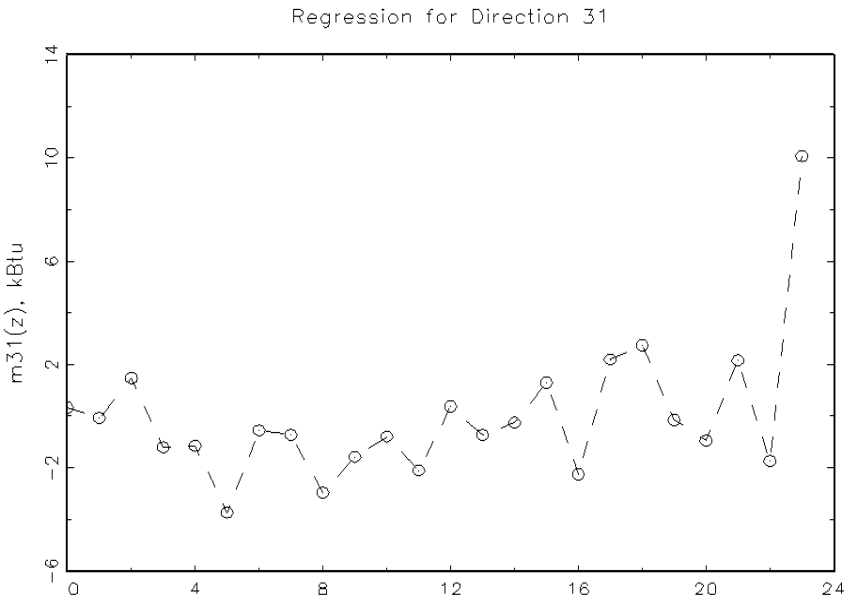

Figure A.31 Total combined income in the past 12 months

0	Less than $2,500
1	$2,500 to $4,999
2	$5,000 to $7,499
3	$7,500 to $9,999
4	$10,000 to $14,999
5	$15,000 to $19,999
6	$20,000 to $24,999
7	$25,000 to $29,999
8	$30,000 to $34,999
9	$35,000 to $39,999
10	$40,000 to $44,999
11	$45,000 to $49,999
12	$50,000 to $54,999
13	$55,000 to $59,999
14	$60,000 to $64,999
15	$65,000 to $69,999
16	$70,000 to $74,999
17	$75,000 to $79,999

18	$80,000 to $84,999
19	$85,000 to $89,999
20	$90,000 to $94,999
21	$95,000 to $99,999
22	$100,000 to $119,999
23	$120,000 or more