Functional forms in energy demand modeling

Jay Zarnikau*

Frontier Associates, 4131 Spicewood Springs Road, Suite O-3, Austin, TX 78759, USA

Abstract

In the estimation of demand functions for energy resources, linear, log–linear and translog functional forms are commonly assumed. It has often been questioned whether such functional forms can indeed accurately represent the underlying relationships between the demand for various energy resources and explanatory variables such as energy prices, weather variables, income and other factors. This paper compares linear, log–linear and translog share equation functional forms against a non-parametric function. Bootstrapping methods are used to test the validity of using the three parametric functional forms in models of residential energy demand. Cross-sectional household-level data from the US BLS Consumption Expenditure survey and other government datasets are used. Each of the parametric functional forms tested performs poorly, suggesting that they may be insufficiently flexible to provide valid results in certain applications.
© 2003 Elsevier Science B.V. All rights reserved.

1. Introduction

Parametric econometric models of energy demand are commonly used to predict the future energy needs under alternative economic and policy scenarios. Elasticity estimates from such models are used to analyze how changes in energy prices, tax changes, income, weather and other factors might affect the demand for various energy resources.[1]

---

*Corresponding author. Tel.: +1-512-372-8778; fax: +1-512-372-8932.

E-mail address: jayz@frontierassoc.com (J. Zarnikau).

[1] In energy demand modeling, there are two general prevalent approaches: econometric modeling and end-use modeling. The end-use or engineering models take a 'bottom-up' approach and estimate electricity demand based on equipment saturations, efficiencies and usage. In the end-use models, relative energy prices may be used as an input into appliance choice modeling. Econometric methods are often used when there is little available information on equipment or appliance stocks.

The functional forms commonly assumed in parametric energy demand models include linear functional forms, log–linear forms and translog models. In the linear models, various explanatory variables are assumed to affect energy demand in a simple linear fashion. In the log–linear models, the dependent and explanatory variables are transformed into natural logarithms and then regressed. Elasticities may then be readily obtained from the estimated coefficients. The linear and log–linear models are sometimes referred to as ad hoc models, to acknowledge their limited theoretical foundation. In contrast, translog models have some basis in microeconomic theory and are popular in the academic literature. The examples of translog models of energy demand may be found in Uri (1982) and Watkins (1992).

A researcher faced with the task of estimating the heating season (winter) household demand for electricity using cross-sectional data and with limited information about the appliance stocks, the housing stock and other end-use factors, might consider the following functional forms:

$$\text{Linear: } \text{KWH}_i = a_E + b_E*\text{PE}_i + b_N*\text{PN}_i + b_I*\text{INC}_i + b_{\text{HDD}}*\text{HDD}_i \tag{1}$$

$$\text{Log–linear: } \log(\text{KWH})_i = a_E + b_E*\log(\text{PE}_i) + b_N*\log(\text{PN}_i) + b_I*\log(\text{INC}_i) + b_{\text{HDD}}*\log(\text{HDD}_i) \tag{2}$$

$$\text{Translog share equation: } \text{SE}_i = a_E + b_E*\text{PE}_i + b_N*\text{PN}_i + b_I*\text{INC}_i + b_{\text{HDD}}*\text{HDD}_i \tag{3}$$

In these equations, KWH_i is the household's total electricity consumption. SE_i is the share of household i's total expenditures on energy resources which were spent on electricity. PE_i represents the price of electricity faced by the household; while PN_i is the price of natural gas, a common alternative energy source for space heating, cooking and water heating. INC_i is the household's income. Also, HDD represents heating degree days, an indicator of weather and space heating needs. The translog model can be estimated as a single electricity share equation, as in Eq. (3), since only two energy resources are being considered and one expenditure share equation must be dropped when a set of share equations is estimated.[2]

2. Specification testing

While the regression test statistics provide some indication of which, if any, of these parametric functional forms are valid, more formal tests can be constructed by comparing the predicted values provided by the parametric specifications against non-parametric alternatives. Specification tests have been developed by Härdle and Mammen (1993) and Zheng (1996). Under this approach, a non-parametric kernel regression is assumed to provide an estimate of the 'true' relationship. Bootstrapping

---

[2] The coefficients on the price terms in the natural gas expenditure share equation can be derived from the b_E and b_N parameters estimated from the electricity share equation, if certain microeconomic relationships are assumed to hold. Typically, when energy demand functions are estimated, the share equations (rather than the cost function) are directly estimated.

techniques are used to construct confidence intervals around the non-parametric regression line. If the parametric regression line falls within these confidence intervals, this would provide empirical support for the parametric model. If the parametric model yields predicted values that consistently fall outside of the confidence band around the non-parametric models, the validity of the parametric functional form is challenged. This approach to specification testing is developed in a more formal manner below.

In its most general form, a regression model may be written as:

$$y = \hat{m}(x) + \varepsilon \qquad (4)$$

where $m(x)$ represents the parametric or non-parametric relationship between the set of explanatory variables x and the dependent variable y. It is assumed that $E(\varepsilon|z,x) = 0$ and $\mathrm{Var}(\varepsilon|x) = \sigma^2(x)$.

In the non-parametric Nadarya–Watson kernel specification, it is merely assumed that $m(°)$ is 'smooth.' In multivariate applications, x is a d-dimensional variable and we have the independent observations $(x_1, y_1),...,(x_n, y_n)$. The multivariate kernel density estimator is:

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_{i1} - x_1}{h_1},..., \frac{x_{ip} - x_p}{h_p} \right) yi}{\sum_{i=1}^{n} K\left(\frac{x_{i1} - x_1}{h_1},..., \frac{x_{ip} - x_p}{h_p} \right)} \qquad (5)$$

where K is a multidimensional kernel function, on d-dimensional arguments. The bandwidth h is actually a vector of bandwidths $h = (h_1,...,h_d)^T$.

A multiplicative or product kernel function can be used to specify the multi-dimensional kernel function $K(u) = k(u_1,...,u_d)$:

$$K(u) = K(u_1) \cdots K(u_p) \qquad (6)$$

Using the non-parametric Nadarya–Watson smoothing technique, a predicted value is estimated for each multidimensional point x.

In order to estimate confidence bounds around the predicted value from the kernel estimation, the wild bootstrap technique is used.[3] The wild bootstrap involves re-sampling from the estimated residuals:

$$\hat{\varepsilon}_i = y_i - \hat{m}_h(x_i) \qquad (7)$$

These residuals are then used to construct an estimator with a distribution that will approximate the original estimator. In the resampling, the residual is drawn from the two-point distribution, which has zero mean, variance equal to the square

---

[3] See Härdle and Marron (1991).

root of the residual and third moment equal to the cube of the residual. These conditions are satisfied by the function $G_i = \gamma\delta_a + (1-\gamma)\delta_b$, where δ_a and δ_b define point measures at a and b. The parameters a, b and γ at each x_i are given by $a = \varepsilon_i(1-5^{0.5})$, $b = \varepsilon_i(1+5^{0.5})/2$ and $\gamma = (5+5^{0.5})/10$.

Following the resampling, new observations are constructed for the dependent variable:

$$\hat{y}_i^* = \hat{m}_g(x_i) + \hat{\varepsilon}_i^* \tag{8}$$

where the new bandwidth g exceeds the original bandwidth h.[4] The bootstrapped data $\{(x_i, y_i^*)\}^n$ $i=1$ are then used to re-estimate the kernel regression model. Numerous replications permit the distribution of m to be estimated, since the distribution of $m_h(x) - m(x)$ is approximated by $\hat{m}_h^*(x) - \hat{m}_g(x)$.

Once the distribution of m is estimated, confidence intervals may be constructed around the predicted values obtained from the non-parametric model. The distance between the parametric fit m_θ and the non-parametric fit m_h can be measured to test the parametric specification against a non-parametric alternative. Blundell and Duncan (1998) reviewed three proposed test statistics and concluded that they each provide similar and reasonable results. In a multivariate setting, the simplest of the proposed test statistics is a simple squared error statistic offered by Äit-Sahaila et al. (2000):

$$\hat{\Gamma} = \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{m}_\theta - \hat{\mu}_h)^2 \right)\pi \tag{9}$$

where π is a weighting parameter that can be used if weighting is desired. This weighting parameter would be useful if a certain range of x values were of greatest interest. A linear transformation of this statistic converges to a normal distribution with mean zero and estimable variance.

3. Objective

This paper tests whether the linear, log–linear or translog parametric functional forms provide reasonable representations of the 'true' functional relationship between the US household-level demand for electricity and various explanatory variables. Data and modeling techniques are described in the following sections.

4. Dataset

Household-level cross-sectional data are used in this test. The total household electricity cost and natural gas cost for the first quarter of 1994 was obtained from the consumer expenditure (CE) survey database compiled by the US bureau of

---

[4] The need for 'oversmoothing' is explained in Härdle and Marron (1991).

labor statistics (US BLS). Information from the CE for approximately 5235 households was downloaded from a web site maintained by U.C.-Berkeley. In addition to total electricity and natural gas costs, the household's state of residence, total income and certain appliance saturation information (whether domestic space heating, water heating and cooking equipment used electricity or natural gas) was extracted from the CE's consumer unit and characteristics and income file (FMLY).

Some additional variables were appended to the CE database. The average price of electricity and natural gas to the residential energy consumers in each state in 1994, was obtained from the US Department of Energy (US DOE) Energy Information Agency's State Energy Price and Expenditure Report database. Weather data (population-weighted heating and cooling degree days for each state) were obtained from the National Climatic Data Center for the first quarter of 1994.

A winter quarter was selected for this analyses, because summer natural gas use can be rather negligible in the residential sector. In the winter, there presumably is greater competition among electricity and natural gas for space heating energy uses. For similar reasons, quarterly data are used, rather than annual data. Relationships among energy consumption and prices in the winter may be very different from those in the summer due to different space conditioning needs. Thus, smaller time intervals provide more meaningful results for this sector with weather-sensitive energy demand.

Particularly for larger states, there is some unavoidable imprecision in the data. The CE public database provides the respondent's state of residence, but no information on the location of the respondent within the state. Consequently, the respondent is assumed to face the statewide average weather and statewide average prices of electricity and natural gas. Unfortunately, different regions within some states may face weather and energy prices that significantly diverge from these statewide averages.

Many of the CE records contained missing values. For example, state identifiers were often deleted from the public database to mask the identities of respondents in sparsely populated states. Occasionally, household income information was missing. Any household record with incomplete information was deleted from the database.

Households that did not consume any natural gas during the quarter were also deleted. Natural gas is not available in many rural areas of the US. Consequently, it was felt that including in this analyses, households that might not actually have natural gas as an option to them might lead to biased results.

Once all observations with missing values or zero natural gas use were deleted, 1341 complete observations remained.

5. Modeling approach

To estimate the parametric models [Eqs. (1)–(3)] and to obtain predicted values from these models, the relationships were estimated using OLS in SAS software.[5]

Since each of the three parametric models use slightly different (in some cases,

---

[5] The SAS MODEL procedure was used.

transformed) variables, three different non-parametric models were constructed, with variables corresponding to their parametric counterparts:

Alternative to linear model:

$$\text{KWH}_i = m(\text{PE}_i, \text{PN}_i, \text{INC}_i, \text{HDD}_i) \tag{10}$$

Alternative to log–linear model:

$$\log(\text{KWH})_i = m(\log(\text{PE}_i), \log(\text{PN}_i), \log(\text{INC}_i), \log(\text{HDD}_i)) \tag{11}$$

Alternative to translog share equation:

$$\text{SE}_i = m(\text{PE}_i, \text{PN}_i, \text{INC}_i, \text{HDD}_i) \tag{12}$$

These multivariate non-parametric regression models were estimated using XploRe software. While 1341 observations were used in the parametric regressions, only the first 950 were used in the non-parametric estimation, due to a size limit in XploRe. A product quartic kernel function was used in the Nadarya–Watson estimation. The software's default bandwidth estimator—20% of the range of x—was used.

Bootstrapping was performed with XploRe's wild bootstrap procedure. One-thousand simulations were performed on Eqs. (10)–(12).[6] The resulting 950 by 1000 matrix of predicted values (the number of observations by the number of simulations), was imported into a SAS program to calculate confidence intervals.

A spreadsheet was used to compare the predicted values from the parametric models to the confidence intervals around the non-parametric predicted values. The XploRe re-orders observations when performing non-parametric regression estimation. Consequently, the predicted values corresponding to the x matrix re-ordered by XploRe and the estimated coefficients from the parametric models were used to re-calculate predicted values for the parametric cases.

6. Results

Some of the variables used in this analyses are graphically depicted in a smoothed form in Figs. 1–4. Electricity consumption per household is presented in Fig. 1. This variable, which serves as the dependent variable in Eq. (1), is distributed non-normally, but has only one major peak. When the electricity consumption is plotted against heating degree days using a Gaussian kernel to achieve smoothing, two peaks are evident (see Fig. 2). Fig. 3 shows two major and one minor peaks when electricity consumption is plotted against the price of electricity (in dollars per MMBtu). Fig. 4 suggests a simple correlation between income and electricity consumption.

---

[6] For each of the three non-parametric models, each bootstrap procedure with 1000 simulations required about 3 h of run time on a personal computer with a 1 GHz processor.

Density with Quadratic Kernel
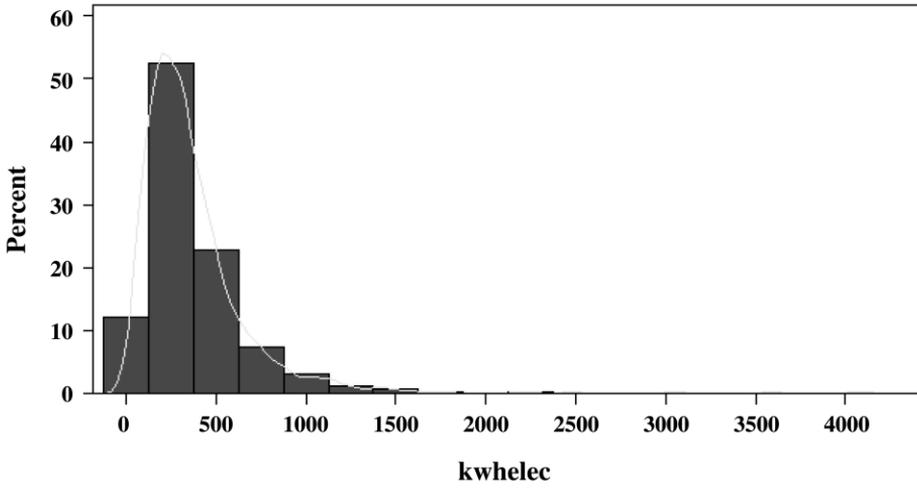
kWh Consumption in First Quarter of 1994

Fig. 1. Density with quadratic kernel: kW h consumption in first quarter of 1994.

Density with Gaussian Kernel

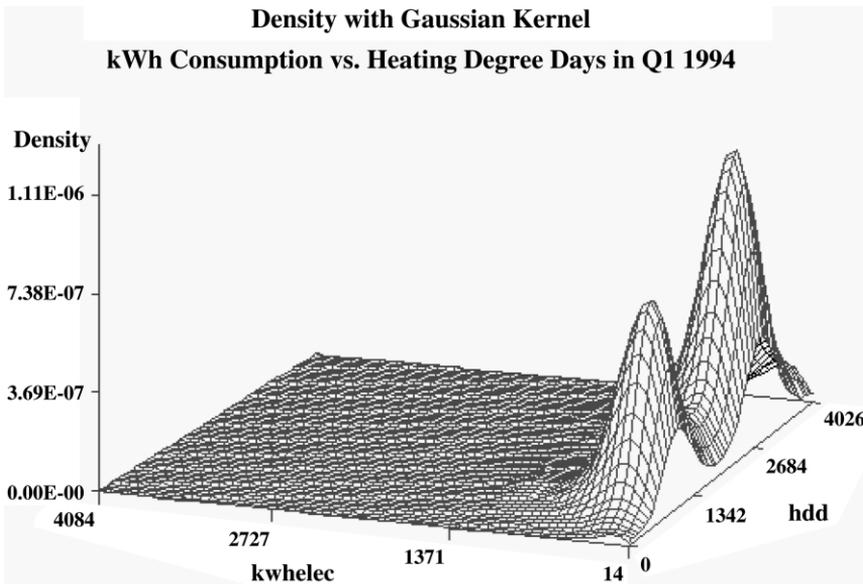**kWh Consumption vs. Heating Degree Days in Q1 1994**

Fig. 2. Density with Gaussian kernel: kW h consumption vs. heating degree days in first quarter of 1994.

Density with Gaussian Kernel
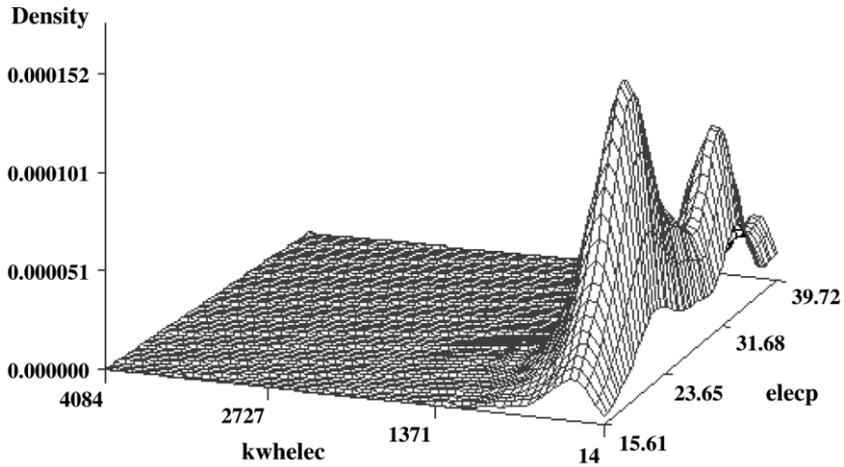**kWh Consumption vs. Electricity Price on Q1 1994**

Fig. 3. Density with gaussian kernel: kW h consumption vs. electricity price in first quarter of 1994.

Density with Gaussian Kernel
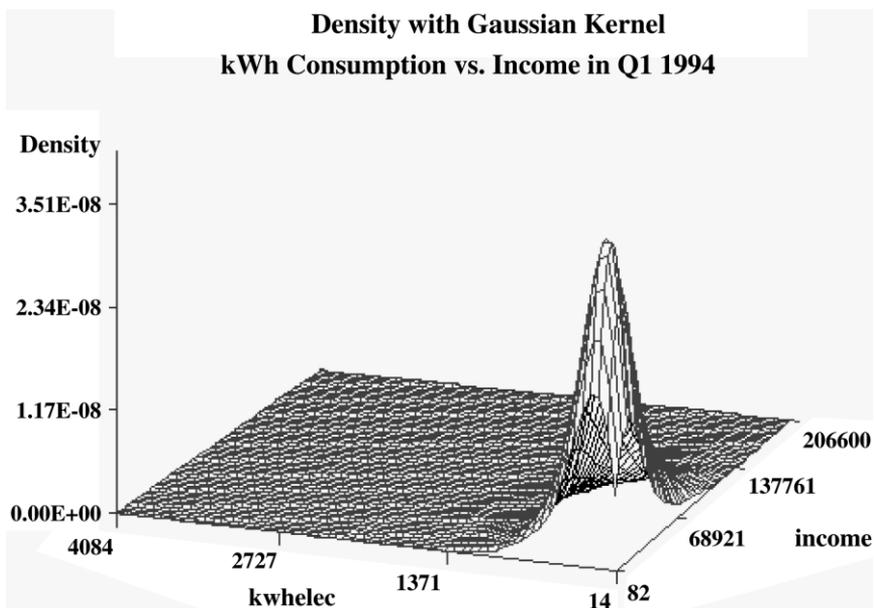**kWh Consumption vs. Income in Q1 1994**

Fig. 4. Density with Gaussian kernel: kW h consumption vs. income in first quarter of 1994.

Table 1
Relationship of parametric predicted values to non-parametric confidence intervals

Model being tested	Within bounds (%)	Below bounds (%)	Above bounds (%)
Linear	29.3	41.9	28.8
Log–linear	17.8	55.6	26.6
Translog	10.3	54.9	34.7

All three of the parametric regressions performed rather poorly. The highest reported R^2 was 0.235, from the translog expenditure share equation. This is not too surprising. Factors other than the four explanatory variables considered here could have a considerable influence upon electricity demand, including the presence of building energy codes in a state or locality, the age of the dwelling, the size of the dwelling and appliance holdings. However, increasing the number of explanatory variables would hamper the non-parametric analyses, as discussed later.

Using the bootstrap simulation results, 95% confidence intervals were constructed around the non-parametric predicted values from each of the three non-parametric models. Table 1 reports the frequency in which the parametric predicted values fell within these intervals. Here, all of the parametric models performed poorly, especially the translog.[7]

Table 2 provides the average (over all 950 observations) of the S.D. of the non-parametric estimates as calculated from the bootstrap simulations. The test statistic from Eq. (9) is also presented (each observation is given equal weight). Each of the test statistics is many S.D. away from zero. This raises concern that the parametric models are inadequate.

7. Summary and conclusions

Under the premise that a non-parametric kernel regression estimator can provide an optimal, or at least superior, representation of the underlying relationships between electricity consumption and a set of four common explanatory variables, three common parametric model specifications were tested and rejected at normal levels of significance. These results suggest that caution should be exercised when making assumptions regarding the functional form of energy demand models. Yet,

Table 2
Test statistic

Model being tested	S.D.	Test statistic $\left(\hat{\Gamma}\right)$
Linear	14	96,127,491,744,084
Log–linear	0.028	0.1493
Translog	0.0073	0.0778

[7] These confidence intervals are difficult to present in graphical form, since there are four independent variables.

some limitations to this non-parametric approach to testing functional forms must be acknowledged.

In one sense, the poor performance of the parametric models should not be too surprising. The goodness-of-fit statistics were poor to begin with. It is difficult to model electricity consumption across numerous US states, since building codes, building practices, the age of the housing stock and appliance holdings can vary considerably across the country. The addition of such variables would no-doubt enhance the parametric analyses, but would hamper the estimation of the non-parametric alternatives, since non-parametric methods require much larger sample sizes as the number of explanatory variables is increased.

One might ask whether the rejection of linear, log–linear and translog functional forms in a cross-sectional context would also imply that these functional forms would be rejected in a (more common) time-series context. Unfortunately, it is difficult to test this question using the methods presented here. Although, a fairly large number of observations were used in these analyses, the sample size was barely adequate to test the validity of these functional forms in a cross-sectional context. In time-series contexts, fewer observations are normally available, and the non-parametric methods discussed here might prove impossible to apply.

Acknowledgments

References

Äit-Sahaila, Y., Bickel, P., Stoker, T., 2000. Goodness-of-fit tests for kernel regression with an application to option-implied volatilities. Princeton University Working Paper.

Blundell, R., Duncan, A., 1998. Kernel regression in empirical microeconomics. J. Hum. Resour. XXXIII, 63–87.

Härdle, W., Mammen, E., 1993. Comparing nonparametric vs. parametric regression fits. Ann. Stat. 21 (4), 1926–1947.

Härdle, W., Marron, J.S., 1991. Bootstrap simultaneous error bars for nonparametric regression. Ann. Stat. 19 (2), 779–796.

Uri, N., 1982. The industrial demand for energy. Resour. Energy 4, 27–57.

Watkins, G.C., 1992. The econometric analysis of energy demand: perspectives of a practitioner. In: Hawdon, D. (Ed.), Energy Demand: Evidence and Expectations. Surrey University Press.

Zheng, J., 1996. A consistent test of functional form via non-parametric estimation functions. J. Econ. 75, 263–290.

Further reading

Cosidine, T., 1989. Separability, functional form and regulatory policy in models of interfuel substitution. Energy Econ. 82–94.

Hall, P., 1992. On bootstrap confidence intervals in nonparametric regression. Ann. Stat. 20 (2), 690–711.

Härdle W., 1989. Applied Nonparametric Regression. Cambridge University Press.

Härdle, W., Klinke, S., Müller, M., XploRe Learning Guide, MD Tech.

Nerlove, M., 1997. Notes on Monte Carlo, bootstrapping, and estimation by simulation.

Qin, S., Damien, P., Zarnikau, J., Mentrup, G., 1998. Assessing separability in production functions: a Bayesian approach. Under review.

Wilson, P., Carey, K., 2000. Non-parametric analysis of returns to scale in the US hospital industry. Draft.

Zarnikau, J., Guermouche, S., Schmidt, P., 1996. Can different energy resources be added or compared? Energy Int. J. 21, 483–491.