Volume Title: Scanner Data and Price Indexes

Volume Author/Editor: Robert C. Feenstra and Matthew D. Shapiro, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-23965-9

Volume URL: http://www.nber.org/books/feen03-1

Conference Date: September 15-16, 2000

Publication Date: January 2003

Title: Hedonic Regressions. A Consumer Theory Approach

Author: W. Erwin Diewert

URL: http://www.nber.org/chapters/c9740

Hedonic Regressions
A Consumer Theory Approach

Erwin Diewert

10.1 Introduction

This paper started out as a comment on Silver and Heravi (chap. 9 in this volume). This very useful and interesting paper follows in the tradition started by Silver (1995), who was the first to use scanner data in a systematic way in order to construct index numbers. In the present paper by Silver and Heravi, the authors collect an enormous data set on virtually all sales of washing machines in the United Kingdom for the twelve months in the year 1998. They use this detailed price and quantity information, along with information on the characteristics of each machine, in order to compute various aggregate monthly price indexes for washing machines, taking into account the problems associated with the changing quality of washing machines. In particular, the authors consider three broad types of approach to the estimation of quality-adjusted prices using scanner data:

- the usual time series dummy variable hedonic regression technique, which does not make use of quantity data on sales of models
- matched model techniques, in which unit values of matched models in each of the two periods being compared are used as the basic prices to go along with the quantities sold in each period (and then ordinary index number theory is used to aggregate up these basic prices and quantities)
- an exact hedonic approach based on the work of Feenstra (1995).

The authors also used their scanner database on washing machines in order to replicate statistical agency sampling techniques.

What I found remarkable about the authors' results is that virtually all[1] of their calculated price indexes showed a very substantial drop in the quality-adjusted prices for washing machines, of about 6 percent to 10 percent over the year. Most of their indexes showed a drop in the aggregate price of washing machines in the 8–10 percent range. In the U.K. Retail Price Index (RPI), washing machines belong to the electrical appliances section, which includes a wide variety of appliances, including irons, toasters, refrigerators, and so on. From January 1998 to December 1998, the electrical appliances RPI component went from 98.6 to 98.0, a drop of 0.6 percentage points. Now it may be that the non-washing machine components of the electrical appliances index increased in price enough over this period to cancel out the large apparent drop in the price of washing machines, but I think that this is somewhat unlikely. Thus we have a bit of a puzzle: why do scanner data and hedonic regression studies of price change find, on average, much smaller increases in price compared to the corresponding official indexes that include the specific commodity being studied?[2] One explanation for this puzzle (if it is a puzzle) might run as follows. At some point in time, the statistical agency initiates a sample of models whose prices are to be collected until the next sample initiation period. Unless some of these models disappear, no other models will be added to the sample. Thus, what may be happening is that the market throws up new models over the period of time between sample initiations. These new models benefit from technical progress and tend to have lower prices (quality adjusted) than the models that the statistical agency is following. In theory, the producers of these outmoded models should drop their prices to match the new competition, but perhaps instead they simply stop producing these outmoded models, leaving their prices unchanged (or not dropping them enough). However, until every last model of these outmoded models is sold, the statistical agency continues to follow their price movements, which are no longer representative of the market.[3] If a model disappears, there is the possibility that the replacement model chosen by the statistical agency is not linked in at a low enough quality-adjusted price,[4] since the use of hedonic regressions is

1. The one exception was a unit value index that was the average price over all washing machines with no adjustments for the changing mix of machines. This quality-unadjusted index showed a drop of only 1 percent over the year. It is particularly interesting that Feenstra's (1995) exact hedonic approach gave much the same answers as the other approaches.

2. See Diewert (1998) for a review of the scanner data studies up to that point in time.

3. If this hypothesis is true, older models should have a tendency to have positive residuals in hedonic regressions. Berndt, Griliches, and Rappaport (1995, 264); Kokoski, Moulton, and Zieschang (1999, 155;); and Koskimäki and Vartia (2001, 4) find evidence to support this hypothesis for desktop computers, fresh vegetables, and computers, respectively.

4. Also when a model disappears, typically statistical agencies ask their price collectors to look for the model that is the closest substitute to the obsolete model, which means that the closest model is also approaching obsolescence.

not very widespread in statistical agencies. These two factors may help to explain why the hedonic regression approach tends to give lower rates of price increase in rapidly changing markets compared to the rates obtained by statistical agencies.

There is another factor that may help to explain why scanner data studies that use matched samples obtain lower rates of price increase (or higher rates of price decrease, as in the case of the washing machines) than those obtained by statistical agencies. Consider the list of models at the sample initiation period. Some of these models will turn out to be "winners" in the marketplace; that is, they offer the most quality-adjusted value.[5] Now, over time, consumers will buy increasing amounts of these winning models, but this in turn will allow the producers of these winning models to lower their prices, since their per unit output fixed costs will be lower as their markets expand. In a scanner data superlative index number computation of the aggregate market price over all models, these "winner" models that have rapid declines in price will get a higher quantity weighting over time, leading to a lower overall measure of price change than that obtained by the statistical agency, since the agency will be aggregating its sample prices using fixed weights.[6]

I do not have any substantial criticisms of the Silver and Heravi paper; I think that they have done a very fine job indeed.

Since I do not have any substantial criticisms of the paper, the question is: what should I do in the remainder of this comment? What I will do is discuss various methodological issues that the authors did not have the space to cover.[7]

Thus, in section 10.2 below, I revisit Sherwin Rosen's (1974) classic paper on hedonics in an attempt to get a much simpler model than the one that he derived. In particular, I make enough simplifying assumptions so that Rosen's very general model reduces down to the usual time series dummy variable hedonic regression model used by Silver and Heravi. The assumptions that are required to get this simple model are quite restrictive, but I hope that, in the future, other researchers will figure out ways of relaxing some of these assumptions. It should be mentioned that I take a traditional consumer demand approach to the problems involved in setting up an

5. These models should have negative residuals at the sample initiation period in a hedonic regression.

6. This point is made by Berndt and Rappaport (2001). However, it is interesting that both Silver and Heravi and Berndt, Griliches, and Rappaport (1995) find that this weighting bias was relatively low in their washing machine and computer studies in which they compared matched model superlative indexes with the results of unweighted hedonic regressions. Berndt, Griliches, and Rappaport found this weighting bias for computers to be around 0.7 percentage points per year.

7. I should mention that many of the methodology questions are discussed more fully in a companion paper that deals with television sets in the United Kingdom rather than washing machines; see Silver (1999b).

econometric framework for estimating hedonic preferences; that is, I do not attempt to model the producer supply side of the market.[8] Another major purpose of this section is to indicate why linear hedonic regression models (where the dependent variable is the model price and the time dummy enters in the regression in a linear fashion) are unlikely to be consistent with microeconomic theory.

In section 10.3, we look at the problems involved in choosing a functional form for the hedonic regression. Some of the issues considered in this section are

- a comparison between the three most commonly used functional forms for hedonic regressions;
- how hedonic regression techniques can be used in order to model the choice of package size;
- whether we should choose flexible functional forms when undertaking hedonic regressions; and
- whether we should use nonparametric functional forms.

Silver and Heravi noted that there is a connection between matched model techniques for making quality adjustments and hedonic regression techniques: essentially, the hedonic method allows information on non-matching observations to be used, whereas information on models that suddenly appear or disappear in the marketplace must be discarded using the matched model methodology. Triplett (2001) has also considered the connection between the two approaches in an excellent survey of the hedonic regression literature. One of the most interesting results that Triplett derives is a set of conditions that will cause a hedonic regression model to give the same results as a matched model. In section 10.4, we generalize this result to cover a more general class of regression models than considered by Triplett, and we extend his results from the two-period case to the many-period case.

One of the features of the Silver and Heravi paper is their use of sales information on models as well as the usual model price and characteristics information that is used in traditional hedonic regression exercises. In section 10.5 below, we look at some of the issues involved in running hedonic regressions when sales information is available.

Section 10.6 provides some comments on Feenstra's (1995) exact hedonic price index approach, which is used by Silver and Heravi. Our tentative conclusion is that it is not really necessary to use Feenstra's approach if one is willing to make the simplifying assumptions that we make in section 10.2 below.

Section 10.7 generalizes our hedonic model presented in section 10.2 to a

---

8. Thus I am following Muellbauer's (1974, 977) example: he says that his "approach is unashamedly one-sided; only the demand side is treated. . . . Its subject matter is therefore rather different from that of the recent paper by Sherwin Rosen. The supply side and the simultaneity problems which may arise are ignored."

more general situation in which completely separate hedonic regressions are run in each period, as opposed to one big hedonic regression run over all periods in the sample.

Section 10.8 concludes.

10.2   The Theory of Hedonic Price Indexes Revisited

Hedonic regression models pragmatically regress the price of one unit of a commodity (a "model" or "box") on a function of the characteristics of the model and a time dummy variable. It is assumed that a sample of model prices can be collected for two or more time periods along with a vector of the associated model characteristics. An interesting theoretical question is whether we can provide a microeconomic interpretation for the function of characteristics on the right hand side of the regression.

Rosen (1974) in his classic paper on hedonics does this. However, his economic model turns out to be extremely complex. In this section, we will rework his model,[9] making two significant changes:

- We will assume that every consumer has the same *separable subutility function,* $f(z_1, \ldots, z_N)$, which gives the consumer the subutility $Z = f(\mathbf{z})$ from the purchase of one unit of the complex hedonic commodity that has the vector of characteristics $\mathbf{z} \equiv (z_1, \ldots, z_N)$.[10]
- The subutility that the consumer gets from consuming Z units of the hedonic commodity is combined with the consumption of X units of a composite "other" commodity to give the consumer an overall utility of $u = U^t(X, Z)$ in period t, where U^t is the period t "macro" utility function. Rosen (1974, 38) normalized the price of X to be unity. We will *not* do this; instead, we will have an explicit period t price, p^t, for one unit of the general consumption commodity X.

We start off by considering the set of X and Z combinations that can yield the consumer's period t utility level, u^t. This is the set $\{(X, Z): U^t(X, Z) = u^t\}$, which of course is the consumer's period t indifference curve over equivalent combinations of the general consumption commodity X and the hedonic commodity Z. Now solve the equation $U^t(X,Z) = u^t$ for X as a function of u^t and Z; that is, we have[11]

---

9. We used Rosen's notation, which was somewhat different from that used by Silver and Heravi.

10. We do not assume that all possible models exist in the marketplace. In fact, we will assume that only a finite set of models exists in each period. However, we do assume that the consumer has preferences over all possible models, where each model is indexed by its vector of characteristics, $\mathbf{z} = (z_1, \ldots, z_N)$. Thus each consumer will prefer a potential model with characteristics vector $\mathbf{z^1} = (z_1^1, \ldots, z_N^1)$ over another potential model with the characteristics vector $\mathbf{z^2} = (z_1^2, \ldots, z_N^2)$ if and only if $f(\mathbf{z^1}) > f(\mathbf{z^2})$.

11. If the period t indifference curve intersects both axes, then $g^t(u^t, Z)$ will only be defined for a range of nonnegative Z up to an upper bound.

(1) $X = g^t(u^t, Z).$

We will assume that this indifference curve slopes downward, and, in fact, we will make the stronger assumption that g^t is differentiable with respect to Z and

(2) $\dfrac{\partial g^t(u^t, Z)}{\partial Z} < 0.$

Let p^t and P^t be the prices for one unit of X and Z respectively in period t. The consumer's period t expenditure minimization problem may be defined as follows:

(3) $\min_{X,Z} [p^t X + P^t Z : X = g^t(u^t, Z)] = \min_Z [p^t g^t(u^t, Z) + P^t Z].$

The first-order necessary condition for Z to solve equation (3) is

(4) $\dfrac{p^t \partial g^t(u^t, Z)}{\partial Z} + P^t = 0.$

Equation (4) can now be rearranged to give the price of the hedonic aggregate P^t as a function of the period t utility level u^t and the price of general consumption p^t:

(5) $P^t = -\dfrac{p^t \partial g^t(u^t, Z)}{\partial Z} > 0$

where the inequality follows from the assumption in equation (2) above. We now interpret the right-hand side of equation (5) as the consumer's period t willingness-to-pay price function $w^t(Z, u^t, p^t)$:

(6) $w^t(Z, u^t, p^t) \equiv -\dfrac{p^t \partial g^t(u^t, Z)}{\partial Z}.$

Thus, as we travel down the consumer's period t indifference curve, for each point (indexed by Z) on this curve, equation (6) gives us the amount of money the consumer would be willing to pay *per unit of Z* in order to stay on the same indifference curve, which is indexed by the utility level u^t.

The period t willingness-to-pay value function v^t can now be defined as the product of the quantity of Z consumed times the corresponding per unit willingness-to-pay price, $w^t(Z, u^t, p^t)$:

(7) $v^t(Z, u^t, p^t) \equiv Z w^t(Z, u^t, p^t) = -\dfrac{Z p^t \partial g^t(u^t, Z)}{\partial Z},$

where the last equality follows using equation (6). The function v^t is the counterpart to Rosen's (1974, 38) value or bid function; it gives us the amount of money the consumer is willing to pay in order to consume Z units.

All of the above algebra has an interpretation that is independent of the

hedonic model; it is simply an exposition of how to derive a willingness-to-pay price and value function using a consumer's preferences defined over two commodities. However, we now assume that the consumer has a separable subutility function, $f(z_1, \ldots, z_N)$, that gives the consumer the subutility $Z = f(\mathbf{z})$ from the purchase of one unit of the complex hedonic commodity[12] that has the vector of characteristics $\mathbf{z} \equiv (z_1, \ldots, z_N)$. Note that we have assumed that the function f is time invariant.[13] We now assume that the consumer's period t utility function is $U^t(X, f(\mathbf{z}))$. The above algebra on willingness to pay is still valid. In particular, our new period t willingness-to-pay price function, for a particular model with characteristics $\mathbf{z} = (z_1, \ldots, z_n)$, is

$$(8) \qquad w^t(f(\mathbf{z}), u^t, p^t) \equiv -\frac{p^t \partial g^t(u^t, f(\mathbf{z}))}{\partial Z} .$$

Our new period t willingness-to-pay value function (which is the amount of money the consumer is willing to pay to have the services of a model with characteristics vector $\mathbf{z}$) is

$$(9) \qquad v^t(f(\mathbf{z}), u^t, p^t) \equiv f(\mathbf{z}) w^t(f(\mathbf{z}), u^t, p^t) = -\frac{f(\mathbf{z}) p^t \partial g^t(u^t, f(\mathbf{z}))}{\partial Z} .$$

Now suppose that there are K^t models available to the consumer in period t, where model k sells at the per unit price of P_k^t and has the vector of characteristics $\mathbf{z_k^t} \equiv (z_{1k}^t, \ldots, z_{Nk}^t)$ for $k = 1, 2, \ldots, K^t$. If the consumer purchases a unit of model k in period t, then we can equate the model price P_k^t to the appropriate willingness-to-pay value defined by equation (9), where $\mathbf{z}$ is replaced by $\mathbf{z_k^t}$; that is, the following equations should hold:

$$(10) \qquad P_k^t = -\frac{f(\mathbf{z_k^t}) \, p^t \partial g^t(u^t, f(\mathbf{z_k^t}))}{\partial Z}; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t.$$

What is the meaning of the separability assumption? Suppose the hedonic commodity is an automobile and suppose that there are only three char-

---

12. If a consumer purchases, say, two units of a model at price P that has characteristics z_1, $\ldots, z_N$, then we can model this situation by introducing an artificial model that sells at price $2P$ and has characteristics $2z_1, \ldots, 2z_N$. Thus the hedonic surface, $Z = f(z)$, consists of only the most efficient models including the artificial models.

13. We do not assume that $f(z)$ is a quasi-concave or concave function of z. In normal consumer demand theory, $f(z)$ can be assumed to be quasi-concave without loss of generality because linear budget constraints and the assumption of perfect divisibility will imply that "effective" indifference curves enclose convex sets. However, as Rosen (1974, 37–38) points out, in the case of hedonic commodities, the various characteristics cannot be untied. Moreover, perfect divisibility cannot be assumed, and not all possible combinations of characteristics will be available on the marketplace. Thus, the usual assumptions made in "normal" consumer demand theory are not satisfied in the hedonic context. Note also that although we placed a smoothness assumption on the macro functions $g^t(u, Z)$—the existence of the partial derivative $\partial g^t(u, Z)/\partial Z$—we do not place any smoothness restrictions on the hedonic subutility function $f(z)$.

acteristics: number of seats in the vehicle, fuel economy, and horsepower. The separability assumption means that the consumer can trade off these three characteristics and determine the utility of any auto with any mix of these three characteristics *independently of his or her other choices* of commodities. In particular, the utility ranking of automobile models is independent of the number of children the consumer might have or what the price of gasoline might be. Obviously, the separability assumption is not likely to be exactly satisfied in the real world, but in order to make our model tractable, we are forced to make this somewhat restrictive assumption.

Another aspect of our model needs some further explanation. We are explicitly assuming that consumers cannot purchase fractional units of each model; they can purchase only a nonnegative integer amount of each model; that is, we are explicitly assuming *indivisibilities* on the supply side of our model. Thus, in each period, there are only a finite number of models of the hedonic commodity available, so that while the consumer is assumed to have continuous preferences over all possible combinations of characteristics $(z_1, \ldots, z_N)$, in each period, only a finite number of isolated models are available on the market.

At this point, we further specialize our model. We assume that every consumer has the same hedonic subutility function[14] $f(\mathbf{z})$ and consumer i has the following linear indifference curve macro utility function in period t:

(11) $g_i^t(u_i^t, Z) \equiv -a^t Z + b_i^t u_i^t; \quad t = 1, \ldots, T; \quad i = 1, \ldots, I$

where a^t and b_i^t are positive constants. Thus for each period t and each consumer i, the period t indifference curve between combinations of X and Z is linear, with the constant slope $-a^t$ being the same for all consumers.[15] However, note that we are allowing this slope to change over time. Now differentiate equation (11) with respect to Z and substitute this partial derivative into equation (10). The resulting equations are[16]

---

14. The sameness assumption is very strong and needs some justification. This assumption is entirely analogous to the assumption that consumers have the same homothetic preferences over, say, food. Although this assumption is not justified for some purposes, for the purpose of constructing a price index for food, it suffices since we are mostly interested in capturing the substitution effects in the aggregate price of food as the relative prices of food components vary. In a similar fashion, we are interested in determining how the "average" consumer values a faster computer speed against more memory; that is, we are primarily interested in hedonic substitution effects.

15. We do not require a linear indifference curve globally but only locally over a certain range of purchases. Alternatively, we can view the linear indifference curve as providing a first-order approximation to a nonlinear indifference curve.

16. Comparing equation (12) with equation (10), it can be seen that the simplifying the assumptions in equation (11) enabled us to get rid of the terms $\partial g^t(u_i^t, f(z_k^t))/\partial Z$, which depend on individual consumer indifference curves between the hedonic commodity and other commodities. If we had individual household data on the consumption of hedonic and other commodities, then we could use normal consumer demand techniques in order to estimate the parameters that characterized these indifference curves.

(12) $\qquad P_k^t = p^t a^t f(\mathbf{z_k^t}); \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t.$

Now define the aggregate price of one unit of Z in period t as[17]

(13) $\qquad\qquad\qquad \rho_t \equiv p^t a^t; \quad t = 1, \ldots, T$

and substitute equation (13) into equation (12) in order to obtain our basic system of hedonic equations:[18]

(14) $\qquad P_k^t = \rho_t f(\mathbf{z_k^t}); \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t.$

Now all we need to do is postulate a functional form for the hedonic subutility function f and add a stochastic specification to equation (14) and we have our basic hedonic regression model. The unknown parameters in f along with the period t hedonic price parameters ρ_t can then be estimated.[19]

It is possible to generalize the above model but get the same model shown in equation (14) if we replace the composite "other" commodity X with $h(\mathbf{x})$, where $\mathbf{x}$ is a consumption vector and h is a linearly homogeneous, increasing, and concave aggregator function. Instead of equation (12), under these new assumptions, we end up with the following equations:

(15) $\qquad P_k^t = c(\mathbf{p^t})a^t f(\mathbf{z_k^t}); \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t,$

where $\mathbf{p^t}$ is now the vector of prices for the $\mathbf{x}$ commodities in period t, and c is the unit cost or expenditure function that is dual to h.[20] Now redefine ρ_t as $c(\mathbf{p^t})a^t$, and we still obtain the basic system of hedonic equation (14).

17. We have switched to subscripts from superscripts in keeping with the conventions for parameters in regression models; that is, the constants ρ_t will be regression parameters in what follows. Note also that ρ_t is the product of the price of the "other" commodity p^t times the period t slope parameter a^t. We need to allow this slope parameter to change over time in order to be able to model the demand for high-technology hedonic commodities, which have been falling in price relative to "other" commodities; that is, we think of a^t as decreasing over time for high-technology commodities.

18. Our basic model ends up being very similar to one of Muellbauer's (1974, 988–89) hedonic models; see in particular his equation (32).

19. It is possible to rework the above theory and give it a producer theory interpretation. The counterpart to the expenditure minimization problem in equation (3) is now the following profit maximization problem: $\max_{X, Z}[P^t Z - w^t X: X = g^t(k^t, Z)]$ where Z is hedonic output and P^t is a period t price for one unit of the hedonic output, w^t is the period t price of a variable input, and X is the quantity used of it, k^t is the period t quantity of a fixed factor (capital, say) and g^t is the firm's factor requirements function. Assuming that $Z = f(z)$, we end up with the following producer theory counterpart to equation (10): $P_k^t = f(z_k^t)\partial g^t(k^t, f(z_k^t))/\partial Z$. The counterpart to the assumption in equation (11) is for firm i, $g_k^t(k_i^t, Z) \equiv a^t Z - b_i^t k_i^t$, and the counterpart to equation (12) becomes $P_k^t = w^t a^t f(z_k^t)$. However, the producer theory model assumptions are not as plausible as the corresponding consumer theory model assumptions. In particular, it is not very likely that each producer will have the same period t aggregate price for a unit of variable input w^t, and it is not very likely that each firm producing in the hedonic market will have the same technology parameter a^t. However, the key assumption that will not generally be satisfied in the producer context is that each *producer is able to produce the entire array of hedonic models,* whereas, in the consumer context, it is quite plausible that each consumer has the possibility of purchasing and consuming each model.

20. Define c as $c(\mathbf{p^t}) \equiv \min_x\{\mathbf{p^t} \cdot \mathbf{x} : h(\mathbf{x}) = 1\}$ where $\mathbf{p^t} \cdot \mathbf{x}$ denotes the inner product between the vectors $\mathbf{p^t}$ and $\mathbf{x}$.

Equation (14) has one property that is likely to be present in more complex and realistic models of consumer choice. This property is that the model prices in period t are *homogeneous of degree one* in the general price level p^t. Thus, if p^t is replaced by λp^t for any $\lambda > 0$ (think of a sudden hyperinflation where λ is large), then equations (12) and (14) imply that the model prices should become λP_k^t. Note that this homogeneity property will *not* hold for the following additive hedonic model:

$$(16) \qquad P_k^t = \rho_t + f(\mathbf{z_k^t}); \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t.$$

Thus, I would lean toward ruling out running hedonic regressions based on the linear model of equation (16) on a priori grounds. Note that hedonic models that take the logarithm of the model price P_k^t as the dependent variable will tend to be consistent with our basic hedonic equation (14), whereas linear models like equation (16) will not be consistent with the normal linear homogeneity properties implied by microeconomic theory.

We turn now to a discussion of some of the problems involved in choosing a functional form for the hedonic subutility function $f(\mathbf{z})$.[21]

10.3   Functional Form Issues

10.3.1   Frequently Used Functional Forms

The three most commonly used functional forms in the hedonic regression literature are the log-log, the semilog, and the linear.[22] We consider each in turn.

In the log-log model, the hedonic aggregator function f is defined in terms of its logarithm as

$$(17) \qquad \ln f(z_1, \ldots, z_N) \equiv \alpha_0 + \sum_{n=1}^{N} \alpha_n \ln z_n,$$

where the α_n is the unknown parameters to be estimated. If we take logarithms of both sides of equation (14), use equation (17), and add error term ε_k^t, we obtain the following hedonic regression model:

$$(18) \quad \ln P_k^t = \beta_t + \alpha_0 + \sum_{n=1}^{N} \alpha_n \ln z_{nk}^t + \varepsilon_k^t; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t,$$

where $\beta_t \equiv \ln \rho_t$ for $t = 1, \ldots, T$. In order to identify all of the parameters, we require a normalization on the β_t and α_0. Typically, we set $\beta_1 = 0$, which is equivalent to $a^1 p^1 = 1$. If we want to impose linear homogeneity (or constant returns to scale) on the hedonic subutility function $f(z)$, we can do this by setting $\sum_{n=1}^{N} \alpha_n = 1$.

---

21. Our discussion draws heavily on Triplett (2001) and Berndt (1991, chap. 4).
22. See Berndt (1991, chap. 4) for historical references to the early use of these functional forms.

In the semilog model, the logarithm of the hedonic function $f(z)$ is defined as

$$(19) \qquad \ln f(z_1, \ldots, z_N) \equiv \alpha_0 + \sum_{n=1}^{N} \alpha_n z_n.$$

If we take logarithms of both sides of equation (14), use equation (18), and add error terms ε_k^t, we obtain the following hedonic regression model:

$$(20) \quad \ln P_k^t = \beta_t + \alpha_0 + \sum_{n=1}^{N} \alpha_n z_{nk}^t + \varepsilon_k^t; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t,$$

where $\beta_t \equiv \ln \rho_t$ for $t = 1, \ldots, T$. Again, in order to identify all of the parameters, we require a normalization on the β_t and α_0, such as $\beta_1 = 0$, which is equivalent to $a^1 p^1 = 1$.

The semilog model has a disadvantage compared to the log-log model: it is not possible to impose constant returns to scale on the semilog hedonic function $f(z)$.[23] However, the semilog model has an advantage compared to the log-log model: the semilog model can deal with situations in which one or more characteristics z_{nk}^t are equal to zero, whereas the log-log model cannot. This is an important consideration if new characteristics come on to the market during the sample period.

In the linear model, the hedonic function $f(z)$ is a simple linear function of the characteristics

$$(21) \qquad f(z_1, \ldots, z_N) \equiv \alpha_0 + \sum_{n=1}^{N} \alpha_n z_n.$$

Substituting equation (21) into equation (14) and adding the error term ε_k^t, we obtain the following hedonic regression model:

$$(22) \qquad P_k^t = \rho_t \left(\alpha_0 + \sum_{n=1}^{N} \alpha_n z_{nk}^t \right) + \varepsilon_k^t; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t.$$

Again, in order to identify all of the parameters, we require a normalization on the ρ_t and α_n, such as $\rho_1 = 0$, which is equivalent to $a^1 p^1 = 1$. Unfortunately, equation (22) is a *nonlinear* regression model, whereas the earlier log-log and semilog models were *linear* regression models. Constant returns to scale on the linear hedonic function can be imposed by setting $\alpha_0 = 0$.

---

23. For some purposes, it is convenient to allow the hedonic utility function to be the type of utility function that is assumed in index number theory, where usually it is assumed that the utility function is homogeneous of degree one, increasing and concave. For example, if we want to use the hedonic framework to model *tied purchases* (i.e., two commodities are sold together at a single price), then the hedonic utility function becomes an ordinary utility function, $f(z_1, z_2)$, where z_1 and z_2 are the quantities of the two commodities that are in the tied package. In this situation, it may be reasonable to assume that f is homogeneous of degree one, in which case the price of a package consisting of z_1 and z_2 unit of the two commodities is $c(p_1, p_2) f(z_1, z_2)$, where $c(p_1, p_2) \equiv \min_{z_1, z_2} \{ p_1 z_1 + p_2 z_2 : f(z_1, z_2) = 1 \}$ is the unit cost function that is dual to the utility function f. There are many other applications in which it would be useful to allow f to be a linearly homogeneous function.

The model shown in equation (22) can also readily deal with the introduction into the marketplace of new characteristics.

It can be seen that none of the three models shown in equations (18), (20), and (22) totally dominates the other two models; each of the three models has at least one advantage over the other two.

Due to the nonlinear form of equation (22), this model has not been estimated very frequently, if at all. However, the following closely related model has been estimated countless times:

$$(23) \quad P_k^t = \rho_t + \alpha_0 + \sum_{n=1}^{N} \alpha_n z_{nk}^t + \varepsilon_k^t; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t.$$

As was indicated in the previous section, the linear model shown in equation (23) is unlikely to be consistent with microeconomic theory, and so we cannot recommend its use.

10.3.2   Hedonic Regressions and the Problem of Package Size

For many commodities, the price declines as the volume purchased increases. How can this phenomenon be modeled using the hedonic regression framework?

Suppose that the vector of characteristics $\mathbf{z} \equiv (z_1, \ldots, z_N)$ is a scalar, so that $N = 1$ and the single characteristic quantity z_1 is the *package size;* that is, it is the quantity of a homogeneous commodity that is contained in the package sold. In this case, it is natural to take the hedonic subutility function $f(z_1)$ to be a continuous monotonically nondecreasing function of one variable with $f(0) = 0$. We drop the subscript 1 in what follows.

A simple specification for $f(z)$ is to let it be a piecewise linear, continuous function or a *linear spline.* In the case of three linear segments, the system of estimating equation (14) would look like the following system after adding errors to equation (14): for $t = 1, \ldots, T$, we have:

$$(24) \quad P_k^t = \rho_t \alpha_1 z_k^t + \varepsilon_k^t \quad \text{if} \quad 0 \leq z_k^t \leq z_1^* = \rho_t[\alpha_1 z_1^* + \alpha_2(z_k^t - z_1^*)] + \varepsilon_k^t$$

$$\text{if} \quad z_1^* \leq z_k^t \leq z_2^* = \rho_t[\alpha_1 z_1^* + \alpha_2(z_2^* - z_1^*) + \alpha_3(z_k^t - z_2^*)] + \varepsilon_k^t \quad \text{if} \quad z_2^* \leq z_k^t.$$

The predetermined package sizes, z_1^* and z_2^*, where we switch from one linear segment to the next, are called break points. The unknown parameters to be estimated are $\rho_1, \ldots, \rho_T, \alpha_1, \alpha_2,$ and α_3. As usual, not all of these parameters can be identified, so it is necessary to impose a normalization such as $\rho_1 = 1$.

There are two difficulties with the system of estimating equations (24):

- The regression is nonlinear in the unknown parameters.
- The estimated coefficients $\alpha_1, \alpha_2,$ and α_3 should be nonnegative.[24] If an

24. Pakes (2001) argues that we should not expect our hedonic regression estimates to satisfy monotonicity restrictions based on the strategic behavior of firms as they introduce new

initial regression yields a negative α_i, then the regression can be rerun, replacing α_i with $(\alpha_i)^2$.

We turn now to a discussion of the flexibility properties of an assumed hedonic subutility function $f(\mathbf{z})$.

10.3.3 Flexibility Issues

In normal consumer demand theory, we usually ask that the functional form for the consumer's utility function (or any of its dual representations) be flexible; that is, we ask that our assumed functional form be able to approximate an arbitrary twice continuously differentiable utility function to the second order.[25] In the hedonic regression literature, this requirement that the functional form for the utility function be flexible has generally not been imposed.[26] For example, the functional forms considered in section 10.3.1 are only capable of providing a linear approximation rather than a quadratic one. The reason why flexible functional forms have not been used in the hedonic literature to a greater extent is probably due to the multicollinearity problem; that is, if we attempt to estimate a hedonic subutility function $f(z)$ that is capable of providing a second-order approximation, then it may have too many unknown parameters to be estimated accurately.[27] Nevertheless, it may be useful to consider the costs and benefits of using alternative flexible functional forms in the hedonic context.

For our first flexible functional form for $f(z)$, consider the following translog functional form (see Christensen, Jorgenson, and Lau 1975), which generalizes our earlier log-log hedonic aggregator function defined by equation (17) above:

$$(25) \qquad \ln f(z_1, \ldots, z_N) \equiv \alpha_0 + \sum_{n=1}^{N} \alpha_n \ln z_n + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij} \ln z_i \ln z_j,$$

where the α_n and the α_{ij} are the unknown parameters to be estimated. If we take logarithms of both sides of equation (14), use equation (25), and add error term ε_k^t, we obtain the following translog hedonic regression model:

$$(26) \qquad \ln P_k^t = \beta_t + \alpha_0 + \sum_{n=1}^{N} \alpha_n \ln z_{nk}^t + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij} \ln z_{ik}^t \ln z_{jk}^t + \varepsilon_k^t;$$

models. However, for credibility reasons, it is likely that statistical agencies will want to impose monotonicity restrictions.

25. See Diewert (1974, 127–33; 1993, 158–64) for examples of flexible functional forms.

26. An exception to this statement is the recent paper by Yu (2001). His discussion is similar to our discussion in many respects and is more general in some respects.

27. The situation in normal consumer demand theory can be more favorable to the accurate estimation of flexible functional forms because we will have an entire *system* of estimating equations in the normal context. Thus, if there are N commodities and price and quantity observations for T periods on H households, we will have $H(N-1)T$ degrees of freedom to work with in the usual systems approach to estimating consumer preferences. In the hedonic regression framework, we have $K^1 + K^2 + \ldots + K^T$ or roughly KT degrees of freedom, where K is the average number of models in each period.

$$\alpha_{ij} = \alpha_{ji}; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t,$$

where $\beta_t \equiv \ln \rho_t$ for $t = 1, \ldots, T$. In order to identify all of the parameters, we require a normalization on the β_t and α_0. Typically, we set $\beta_1 = 0$, which is equivalent to $a^1 p^1 = 1$. If we want to impose linear homogeneity (or constant returns to scale) on the hedonic subutility function $f(z)$, we can do this by setting $\sum_{n=1}^{N} \alpha_n = 1$ and imposing the restrictions $\sum_{j=1}^{N} \alpha_{ij} = 0$ for $i = 1, \ldots, N$. Obviously, the translog model shown in equation (26) contains the log-log model shown in equation (18) as a special case.[28]

The translog hedonic model shown in equation (26) has two nice properties:

- The right-hand side of equation (26) is linear in the unknown parameters so that linear regression techniques can be used in order to estimate the unknown parameters.
- Constant returns to scale can readily be imposed on the translog hedonic utility function $f(z)$ without destroying the flexibility of the functional form.

The main disadvantage of the translog hedonic model is that, like the log-log model, it cannot deal with the zero characteristics problem.

For our second flexible functional form, consider the following generalization of the semilog hedonic utility function in equation (19):

$$(27) \qquad \ln f(z_1, \ldots, z_N) \equiv \alpha_0 + \sum_{n=1}^{N} \alpha_n z_n + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij} z_i z_j$$

where the α_n and the α_{ij} are the unknown parameters to be estimated. If we take logarithms of both sides of equation (14), use equation (27), and add error term ε_k^t, we obtain the following semilog quadratic hedonic regression model:

$$(28) \qquad \ln P_k^t = \beta_t + \alpha_0 + \sum_{n=1}^{N} \alpha_n z_{nk}^t + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij} z_{ik}^t z_{jk}^t + \varepsilon_k^t;$$

$$t = 1, \ldots, T; \quad k = 1, \ldots, K^t$$

where $\beta_t \equiv \ln \rho_t$ for $t = 1, \ldots, T$. Again, in order to identify all of the parameters, we require a normalization on the β_t and α_0, such as $\beta_1 = 0$, which is equivalent to $a^1 p^1 = 1$.

The semilog quadratic model has a disadvantage compared to the translog model: it is not possible to impose constant returns to scale on the semilog quadratic hedonic function $f(z)$. Both models share the advantage of being linear in the unknown parameters. However, the semilog quadratic model has an advantage compared to the translog model: the semilog model

---

28. In view of our discussion in section 10.2, the translog $f(z)$ does not have to satisfy any curvature conditions.

can deal with situations in which one or more characteristics z_{nk}^t are equal to zero, whereas the translog model cannot. This is an important consideration if new characteristics come on to the market during the sample period.

For our third flexible functional form for the hedonic utility function $f(z)$, consider the following generalized linear functional form (see Diewert 1971).

$$(29) \qquad f(z_1, \ldots, z_N) \equiv \alpha_0 + \sum_{n=1}^{N} \alpha_n(z_n)^{1/2} + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij}(z_i)^{1/2}(z_j)^{1/2}$$

where the α_n and the α_{ij} are the unknown parameters to be estimated. Note that equation (29) generalizes our earlier linear functional form shown in equation (21).[29] Substituting equation (29) into equation (14) and adding the error term ε_k^t, we obtain the following generalized linear hedonic regression model:

$$(30) \qquad P_k^t = \rho_t[\alpha_0 + \sum_{n=1}^{N} \alpha_n(z_{nk}^t)^{1/2} + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij}(z_{ik}^t)^{1/2}(z_{jk}^t)^{1/2}] + \varepsilon_k^t;$$

$$t = 1, \ldots, T; \quad k = 1, \ldots, K^t.$$

As usual, in order to identify all of the parameters, we require a normalization on the ρ_t, α_n, and α_{ij} such as $\rho_1 = 0$, which is equivalent to $a^1 p^1 = 1$. Unfortunately, equation (30) is a nonlinear regression model, whereas the earlier translog and semilog quadratic models were linear regression models. Constant returns to scale on the generalized linear hedonic function can be imposed by setting $\alpha_n = 0$ for $n = 0,1, \ldots, N$. The model in equation (22) can also readily deal with the introduction into the marketplace of new characteristics.

As was the case in section 10.3.1, none of the three flexible hedonic regression models presented in this section totally dominates the remaining two models. Equations (26) and (28) have the advantage of being linear regression models, whereas equation (30) is nonlinear. Equation (26) cannot deal very well with the introduction of new characteristics during the sample period, whereas equations (28) and (30) can. Constant returns to scale in characteristics can readily be imposed in equations (26) and (30), whereas this is not possible with equation (28). Thus each of the three models has two favorable characteristics and one unfavorable characteristic.

10.3.4 Nonparametric Functional Forms

It is possible to address the functional form problem in a nonparametric manner using generalized dummy variable techniques.[30]

---

29. Let the α_n and α_{ij} for $i \neq j$ all equal 0 in equation (29) and we obtain equation (21).

30. The material that we are going to present in this section is essentially equivalent to what statisticians call an analysis of variance model (a two-way layout with interaction terms); see chapter 4 in Scheffé (1959).

Suppose that there are only two characteristics that are important for the models on the market during periods $t = 1, \ldots, T$. Suppose further that there are only I configurations of the first characteristic and J configurations of the second characteristic during the sample period, where I and J are integers greater than 1.[31] Suppose further that in period t we have K_{ij}^t observations that have first characteristic in group i and second characteristic in group j. Denote the kth observation in period t in this i, j grouping as $z_{ijk}^t = (z_{1ijk}^t, z_{2ijk}^t)$. For this configuration of characteristics, we define the corresponding hedonic utility as follows:

$$(31) \quad f(z_{ijk}^t) \equiv \alpha_{ij}; \quad t = 1, \ldots, T; \ i = 1, \ldots, I; \ j = 1, \ldots, J; \ k = 1, \ldots, K_{ij}^t.$$

Let P_{ijk}^t denote the period t price for observation k that has model characteristics that put it in the i, j grouping of models. Substituting equation (31) into equation (14) and adding the error term ε_{ijk}^t leads to the following (nonlinear) generalized dummy variable hedonic regression model:

$$(32) \qquad\qquad P_{ijk}^t = \rho_t \alpha_{ij} + \varepsilon_{ijk}^t;$$

$$t = 1, \ldots, T; \quad i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K_{ij}^t.$$

As usual, not all of the parameters ρ_t for $t = 1, \ldots, T$ and α_{ij} for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ can be identified and so it is necessary to impose a normalization on the parameters like $\rho_1 = 1$.

The hedonic regression model shown in equation (32) is nonlinear. However, in this case, we can reparameterize our theoretical model so that we end up with a linear regression model. Suppose that we take logarithms of both sides of equation (31). Then, defining $\ln \alpha_{ij}$ as γ_{ij}, we have

$$(33) \qquad\qquad \ln f(z_{ijk}^t) \equiv \gamma_{ij};$$

$$t = 1, \ldots, T; \quad i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K_{ij}^t.$$

Substituting equation (33) into equation (14) after taking logarithms of both sides of equation (14) and adding the error term ε_{ijk}^t leads to the following linear generalized dummy variable hedonic regression model:

$$(34) \qquad\qquad \ln P_{ijk}^t = \beta_t + \gamma_{ij} + \varepsilon_{ijk}^t;$$

$$t = 1, \ldots, T; \quad i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K_{ij}^t,$$

where $\beta_t \equiv \ln \rho_t$ for $t = 1, \ldots, T$. As usual, not all of the parameters β_t for $t = 1, \ldots, T$ and γ_{ij} for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ can be identified, and

---

31. Alternatively, we *group* observations so that all models having a quantity z_1 of the first characteristic between 0 and z_1^* are in group 1, all models having a quantity z_1 of the first characteristic between z_1^* and z_2^* are in group 2, $\ldots$, and all models having a quantity z_1 of the first characteristic between z_{I-1}^* and z_I^* are in group I. We do a similar grouping of the models for the second characteristic. Thus any model k in each period falls into one of IJ discrete groupings of models.

so it is necessary to impose a normalization on the parameters like $\beta_1 = 0$, which corresponds to $\rho_1 = 1$.

Which of the two generalized dummy variable hedonic regression model equations (32) or (34) is "better"? Obviously, they both have exactly the same economic content, but of course, the stochastic specifications for the two models differ. Hence, we would have to look at the statistical properties of the residuals in the two models to determine which is better.[32] However, without looking at residuals, the linear regression model equation (34) will be much easier to implement than the nonlinear model equation (32), especially for large data sets.

The linear generalized dummy variable hedonic regression model equations (32) and (34) have two major advantages over the traditional flexible functional form models listed in section 10.3.3:

- The dummy variable models shown in equations (32) and (34) are completely nonparametric and hence are much more flexible than traditional flexible functional forms.
- The dummy variable models can easily accommodate discrete characteristic spaces.

However, the dummy variable hedonic regressions also have some disadvantages:

- There can be an enormous number of parameters to estimate, particularly if there are a large number of distinct characteristics.
- If we attempt to reduce the number of parameters by having fewer class intervals for each characteristic, we will introduce more variance into our estimated coefficients.
- Different investigators will choose differing numbers of classification cells; that is, differing dummy variable hedonic specifications made by different hedonic operators will choose differing Is and Js, leading to a lack of reproducibility in the models.[33]
- If j is held constant, then the α_{ij} and γ_{ij} coefficients should increase (or at least not decrease) as i increases from 1 to I.[34] Similarly, if i is held constant, then the α_{ij} and γ_{ij} coefficients should increase (or at least not decrease) as j increases from 1 to J. The regression model equations

---

32. There is another consideration involved in choosing between equations (32) and (34). The parameters that we are most interested in are the ρ_t, not their logarithms, the β_t. However, as Berndt (1991, 127) noted, "explaining variations in the natural logarithm of price is not the same as explaining variations in price." Thus, Silver and Heravi and Triplett (2001) both note that the antilog of the least squares estimator for β_t will not be an unbiased estimator of ρ_t under the usual stochastic specification, and they cite Goldberger (1968) for a method of correcting this bias. Koskimäki and Vartia (2001, 15) also deal with this problem. These considerations would lead one to favor estimating equation (32) rather than equation (34).

33. The reproducibility issue is very important for statistical agencies.

34. We follow the usual convention that individual characteristics are defined in such a way that a larger quantity of any characteristic yields a larger utility to the consumer.

(32) and (34) ignore these restrictions, and it may be difficult to impose them.[35]

Nevertheless, I believe that these generalized dummy variable hedonic regression techniques look very promising. These models, along with other nonparametric models, deserve a serious look by applied researchers.

10.4   Hedonic Regressions and Traditional Methods for Quality Adjustment

Silver and Heravi demonstrated how traditional matched model techniques for making quality adjustments can be reinterpreted in the context of hedonic regression models. Triplett (2001) and Koskimäki and Vartia (2001, 9) also have some results along these lines. In this section, we review two of Triplett's results.

Suppose that the hedonic regression equation (14) holds in period t and we want to compare the quality of model 1 with that of model 2. Then it can be seen that the first two of equations (14) imply that the utility of variety 2 relative to variety 1 is

$$(35) \qquad \frac{f(\mathbf{z_2^t})}{f(\mathbf{z_1^t})} = \frac{(P_2^t/\rho_t)}{(P_1^t/\rho_t)} = \frac{P_2^t}{P_1^t};$$

that is, the utility or intrinsic value to the consumer of model 2 relative to the utility of model 1 is just the price ratio, P_2^t/P_1^t. Thus, in this case, a quality adjustment that falls out of a hedonic regression model is equivalent to a "traditional" statistical agency quality adjustment technique, which is to use the *observed price ratio* of the two commodities in the same period as an indicator of the relative quality of the two commodities.[36]

In a second example showing how traditional statistical agency quality adjustment techniques can be related to hedonic regressions, Triplett (2001) showed that under certain conditions, the usual matched model method for calculating an overall measure of price change going from one period to the next (using geometric means) was identical to the results obtained using a hedonic regression model.[37] We now look at Triplett's result in a somewhat more general framework.

Recall our standard hedonic regression model equation (14) above. Suppose further that the logarithm of $f(z)$ is a linear function in J unknown parameters, $\alpha_1, \ldots, \alpha_J$; that is, we have

$$(36) \qquad \ln f(\mathbf{z_k^t}) \equiv \alpha_1 + \sum_{j=2}^{J} x_j(\mathbf{z_k^t})\, \alpha_j; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t$$

---

35. Note that there are comparable monotonicity restrictions that the continuous hedonic models listed in sections 10.3.1 and 10.3.3 should also satisfy, and it will be difficult to impose these conditions for these models as well.

36. We are ignoring the error terms in the hedonic regressions in making this argument.

37. Koskimäki and Vartia (2001, 9) state a similar more general result, which is very similar to the result that we obtain below.

where the functions $x_j(z_k^t)$ are known. Note that we have assumed that $x_1(\mathbf{z_k^t}) \equiv 1$; that is, we have assumed that the functional form for $\ln f(z)$ has a constant term in it. Now take logarithms of both sides of equation (14), substitute equation (36) into these logged equations, and add the stochastic term ε_k^t to obtain the following system of regression equations:

$$(37) \quad \ln P_k^t = \beta_t + \alpha_1 + \sum_{j=2}^{J} x_j(z_k^t)\alpha_j + \varepsilon_k^t; \quad t = 1, \ldots, T; \quad k = 1, \ldots, K^t$$

where, as usual, we have defined $\beta_t \equiv \ln \rho_t$ for $t = 1, \ldots, T$. A normalization is required in order to identify all of the parameters in equation (37). We choose the normalization $\rho_1 = 1$, which translates into the following normalization:

$$(38) \qquad\qquad\qquad \beta_1 = 0.$$

Using matrix notation, we can write the period t equations in equation (37) as

$$(39) \qquad\qquad \mathbf{y^t} = \mathbf{1^t}\,\beta_t + \mathbf{X^t}\,\boldsymbol{\alpha} + \boldsymbol{\varepsilon}^t; \quad t = 1, \ldots, T$$

where $\mathbf{y^t} \equiv [\ln P_1^t, \ldots, \ln P_{K_t}^t]'$ is a period t vector of logarithms of model prices (where $'$ denotes the transpose of the preceding vector), β_t is the scalar parameter $\ln \rho_t$, $\mathbf{1^t}$ is a column vector consisting of K^t ones, $\mathbf{X^t}$ is a K^t by J matrix of exogenous variables, $\boldsymbol{\alpha} \equiv [\alpha_1, \ldots, \alpha_J]'$ is a column vector of parameters that determine the hedonic subutility function, and $\boldsymbol{\varepsilon}^t \equiv [\varepsilon_1^t, \ldots, \varepsilon_{K_t}^t]'$ is a column vector of period t disturbances. Now rewrite the system of equations (39) in stacked form as

$$(40) \qquad\qquad\qquad \mathbf{y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}' \equiv [y^{1'}, \ldots, y^{T'}]$, $\boldsymbol{\varepsilon}' \equiv [\varepsilon^{1'}, \ldots, \varepsilon^{T'}]$, $\boldsymbol{\gamma}' \equiv [\beta_2, \beta_3, \ldots, \beta_T, \alpha_1, \ldots, \alpha_J]$, and the matrix $\mathbf{W}$ is a somewhat complicated matrix that is constructed using the column vectors $\mathbf{1^t}$ and the K^t by J matrices $\mathbf{X^t}$ for $t = 1, \ldots, T$.[38]

The vector of least squares estimators for the components of $\boldsymbol{\gamma}$ is

$$(41) \qquad\qquad\qquad \boldsymbol{\gamma}^* \equiv (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}.$$

Define the vector of least squares residuals $\mathbf{e}$ by

$$(42) \qquad\qquad \mathbf{e} \equiv \mathbf{y} - \mathbf{W}\boldsymbol{\gamma}^* = \mathbf{y} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}.$$

It is well known that the vector of least squares residuals $\mathbf{e}$ is orthogonal to the columns of $\mathbf{W}$; that is, we have

$$(43) \qquad \mathbf{W}'\mathbf{e} = \mathbf{W}'[\mathbf{y} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}] = \mathbf{W}'\mathbf{y} - \mathbf{W}'\mathbf{y} = \mathbf{0}'_{T-1+J}$$

where $\mathbf{0}_{T-1+J}$ is a vector of zeros of dimension $T - 1 + J$. Now premultiply both sides of $\mathbf{e} \equiv \mathbf{y} - \mathbf{W}\boldsymbol{\gamma}^*$ by the transposes of the first $T - 1$ columns of $\mathbf{W}$. Using equation (43), we obtain the following equations:

38. Note that we used the normalization shown in equation (38) in order to eliminate the parameter β_1 from the parameter vector $\boldsymbol{\gamma}$.

(44) $0 = \mathbf{1}^{t\prime}\mathbf{y}^t - \mathbf{1}^{t\prime}\mathbf{1}^t\beta_t^* - \mathbf{1}^{t\prime}\mathbf{X}^t\boldsymbol{\alpha}^*; \quad t = 2, 3, \ldots, T$

where β_t^* is the least squares estimator for β_t and $\boldsymbol{\alpha}^* \equiv [\alpha_1^*, \ldots, \alpha_J^*]'$ is the vector of least squares estimators for $\boldsymbol{\alpha} \equiv [\alpha_1, \ldots, \alpha_J]'$. Now column T in **W** corresponds to the constant term α_1 and hence is a vector of ones. Premultiply both sides of equation (42) by this column, and by using equation (43), we obtain the following equation:

(45) $0 = \sum_{t=1}^{T} \mathbf{1}^{t\prime}\mathbf{y}^t - \sum_{t=2}^{T} \mathbf{1}^{t\prime}\mathbf{1}^t\beta_t^* - \sum_{t=2}^{T} \mathbf{1}^{t\prime}\mathbf{X}^t\boldsymbol{\alpha}^*.$

Substitute equation (44) into equation (45) in order to obtain the following equation:

(46) $\mathbf{1}^{1\prime}\mathbf{y}^1 = \mathbf{1}^{1\prime}\mathbf{X}^1\boldsymbol{\alpha}^*.$

Noting that $\mathbf{1}^{t\prime}\mathbf{1}^t = K^t$ (the number of model prices collected in period t), we can rewrite equation (44) as follows:

(47) $\beta_t^* = \dfrac{1}{K^t} \sum_{k=1}^{Kt} y_k^t - \dfrac{1}{K^t} \mathbf{1}^{t\prime}\mathbf{X}^t\boldsymbol{\alpha}^*; \quad t = 2, 3, \ldots, T.$

The β_t^* defined by the right-hand side of equation (47) can be given an interesting interpretation as an arithmetic average of the vector of quality-adjusted period t logarithmic prices $y^t - X^t\alpha^*$. However, a very interesting result emerges from using equations (46) and (47) if we assume that the sample of model prices is *matched* for all T periods (so that in each period, exactly the same models are priced). If the sample is matched, then each $\mathbf{X}^t$ matrix is exactly the same (and all K^t equal a common sample size K). If the common $\mathbf{X}^t$ matrix is the K by $T - 1 + J$ matrix **X**, then using equations (46) and (47) gives us the following formula for β_t^*:

(48) $\beta_t^* = \dfrac{1}{K} \sum_{k=1}^{K} y_k^t - \dfrac{1}{K} \sum_{k=1}^{K} y_k^1; \quad t = 2, 3, \ldots, T.$

Thus, in the matched sample case, taking the exponential of β_t^* as our estimator of ρ_t and recalling that $y_k^t \equiv \ln P_k^t$, we have

(49) $\rho_t^* \equiv \dfrac{\left(\prod_{k=1}^{K} P_k^t\right)^{1/K}}{\left(\prod_{k=1}^{K} P_k^1\right)^{1/K}} = \left[\prod_{k=1}^{K}\left(\dfrac{P_k^t}{P_k^1}\right)\right]^{1/K}; \quad t = 2, 3, \ldots, T;$

that is, the hedonic regression approach in the matched model case gives exactly the same result for the overall measure of price change going from period 1 to t as what we would get by taking the geometric mean of the matched model price relatives for the two periods under consideration. Triplett indicated that this result was true for the case $T = 2$ and assuming that f was the log-log hedonic utility function described in section 10.3.1.

I think that the Silver and Heravi paper and the Triplett (2001) manual are both very useful in that they indicate very explicitly that traditional matched model techniques for quality adjustment can be quite closely related to the results of a hedonic regression approach. This correspondence between the two methods should help to demystify hedonic methods to some extent. Furthermore, as stressed by Silver and Heravi and Triplett, the statistical advantage in using the hedonic regression approach over the matched model approach increases as the lack of matching increases; that is, the hedonic technique uses all of the model information between the two periods under consideration, whereas the matched model approach can by definition use only the information on models that are present in the marketplace during both periods.

10.5   Hedonic Regressions and the Use of Quantity Weights

The hedonic regression study by Silver and Heravi is relatively unusual in that they not only had data on the prices and characteristics of washing machines sold in the United Kingdom in 1998, but they also had data on the sales of each model. The question that we want to address in this section is: how exactly should quantity data be used in a hedonic regression study?

We start out by considering a very simple model in which there is only one variety in the market during period t, but we have K price observations, P_k^t, on this model during period t, along with the corresponding quantity sold at each of these prices, q_k^t. Under these assumptions, our basic hedonic regression equation (14) for period t become

(50) $$P_k^t = \rho_t f(\mathbf{z_k^t}) = \rho_t; \quad k = 1, 2, \ldots, K$$

where we can set $f(\mathbf{z_k^t}) = 1$, since all K transactions are on exactly the same model.

From viewing equation (50), we see that ρ_t can be interpreted as some sort of average of the K period t observed transaction prices, P_k^t. The *relative frequency* at which the price P_k^t is observed in the marketplace during period t can be defined as

(51) $$\theta_k^t \equiv \frac{q_k^t}{\sum_{i=1}^{K} q_i^t}.$$

The expected value of the discrete distribution of period t prices is

(52) $$\rho_t^* \equiv \sum_{k=1}^{K} \theta_k^t P_k^t = \frac{\sum_{k=1}^{K} q_k^t P_k^t}{\sum_{i=1}^{K} q_k^t} \quad \text{using equation (51).}$$

Note that the far right-hand side of equation (52) is a unit value. Thus quantity data on the sales of a model can be used to form a *representative*

average price for the model in a period, and that representative price is an overall sales weighted average price for the model or a unit value.[39]

How can we derive the unit value estimator for the representative period t price ρ_t using a hedonic regression? There are at least two ways of doing this.

Look at equation k in the system of price equations (50). Since there are q_k^t sales at this price in period t, we could repeat the equation $P_k^t = \rho_t$ a number of times, q_k^t times to be exact. Let $\mathbf{1_k}$ be a vector of dimension q_k^t. Then, using vector notation, we could write rewrite the system of equations (50), repeating each price P_k^t the appropriate number of times that a transaction took place in period t at that price, as follows:

$$(53) \qquad 1_k P_k^t = 1_k \rho_t; \quad k = 1, 2, \ldots, K.$$

Now add error terms to each of equations (53) and calculate the least squares estimator for the resulting linear regression. This estimator turns out to be the unit value estimator ρ_t^* defined by equation (52).

The second way of deriving the unit value estimator for the representative period t price ρ_t using a hedonic regression is to multiply both sides of equation k in equations (50) by the square root of the quantity of model k sold in period t, $(q_k^t)^{1/2}$ and then add an error term, ε_k^t. We obtain the following system of equations:

$$(54) \qquad (q_k^t)^{1/2} P_k^t = (q_k^t)^{1/2} \rho_t + \varepsilon_k^t; \quad k = 1, 2, \ldots, K.$$

Note that the left-hand side variables in equation (54) are known. Now treat equation (54) as a linear regression with the unknown parameter ρ_t to be estimated. It can be verified that the least squares estimator for ρ_t is the unit value estimator ρ_t^* defined by equation (52).[40] Thus we can use a weighted least squares hedonic regression as a way of obtaining a more representative average model price for period t.

The above discussion may help to explain why Silver and Heravi used sales-weighted hedonic regressions in their regression models. The use of quantity-weighted regressions will diminish the influence of unrepresenta-

---

39. One could think of other ways of weighting the prices P_k^t. For example, we could use the expenditure share for all models sold at the price P_k^t during period t equal to $s_k^t \equiv P_k^t q_k^t / \sum_{i=1}^{K} P_i^t q_i^t$ for $k = 1, \ldots, K$ as a weighting factor for P_k^t. The representative period t average price using these weights becomes $\rho_t^{**} \equiv \sum_{k=1}^{K} s_k^t P_k^t$. Note that if we divide this price into the value of period t transactions, $\sum_{i=1}^{K} P_i^t q_i^t$, we obtain the corresponding quantity estimator, $(\sum_{i=1}^{K} P_i^t q_i^{t2} / \sum_{k=1}^{K} (P_k^t)^2 q_k^t$, which is not easy to interpret. On the other hand, if we divide the unit value estimator of aggregate period t price, ρ_t^* defined by equation (53), into the value of period t transactions, $\sum_{i=1}^{K} P_i^t q_i^t$, we obtain the simple sum of quantities transacted during period t, $\sum_{k=1}^{K} q_k^t$, as the corresponding quantity estimator. The use of unit values to aggregate over transactions pertaining to a homogeneous commodity within a period to obtain a single representative price and quantity for the period under consideration was advocated by Walsh (1901, 96; 1921, 88), Davies (1924, 187), and Diewert (1995, 20–24).

40. Berndt (1991, 127) presents a similar econometric argument justifying the weighted least squares model in equation (54) in terms of a model involving heteroskedastic variances for the untransformed model.

tive prices[41] and should lead to a better measure of central tendency for the distribution of quality-adjusted model prices; that is, the use of quantity weights should lead to more accurate estimates of the ρ_t parameters in equation (14).

10.6   Exact Hedonic Indexes

Silver and Heravi spend a considerable amount of effort in evaluating two of Feenstra's (1995) bounds to an exact hedonic index. In section 10.2, we made some rather strong simplifying assumptions on the structure of consumer preferences, assumptions that were rather different from those made by Feenstra. In this section, we look at the implications of our assumptions for constructing exact hedonic indexes.[42]

Recall our basic hedonic equation (14) again: $P_k^t = \rho_t f(\mathbf{z_k^t})$ for $t = 1, \ldots,$ T and $k = 1, \ldots, K^t$. We assume that the price P_k^t is the average price for all the models of type k sold in period t, and we let q_k^t be the number of units sold of model k in period t. Recall that the number of models in the marketplace during period t was K^t.

In this section, we will assume that there are K models in the marketplace over all T periods in our sample period. If a particular model k is not sold at all during period t, then we will assume that P_k^t and q_k^t are both zero. With these conventions in mind, the total value of consumer purchases during period t is equal to

$$(55) \qquad \sum_{k=1}^{K} P_k^t q_k^t = \sum_{k=1}^{K} \rho_t f(\mathbf{z_k}) q_k^t; \quad t = 1, \ldots, T.$$

The hedonic subutility function f has done all of the hard work in our model in converting the utility yielded by model k in period t into a "standard" utility $f(\mathbf{z_k})$ that is cardinally comparable across models. Then, for each model type k, we just multiply by the total number of units sold in period t, q_k^t, in order to obtain the total period t market quantity of the hedonic commodity, Q_t, say. Thus we have[43]

$$(56) \qquad Q_t \equiv \sum_{k=1}^{K} f(\mathbf{z_k}) q_k^t; \quad t = 1, \ldots, T.$$

The corresponding aggregate price for the hedonic commodity is ρ_t. Thus, in our highly simplified model, the aggregate exact period t price and

---

41. Griliches (1961; 1971, 5) made this observation many years ago.

42. Our assumptions are also quite different from those made by Fixler and Zieschang (1992), who took yet another approach to the construction of exact hedonic indexes.

43. This is a counterpart to the quantity index defined by Muellbauer (1974, 988) in one of his hedonic models; see his equation (30). Of course, treating ρ_t as a price for the hedonic commodity quantity aggregate defined by equation (57) can be justified by appealing to Hicks's (1946, 312–13) Aggregation Theorem, since the model prices $P_k^t = \rho_t f(z_k)$ all have the common factor of proportionality, ρ_t.

quantity for the hedonic commodity is ρ_t and Q_t defined by equation (56), which can readily be calculated, provided we have estimated the parameters in the hedonic regression equation (14) and provided that we have data on quantities sold during each period, the q_k^t.[44]

Once ρ_t and Q_t have been determined for $t = 1, \ldots, T$, then these aggregate price and quantity estimates for the hedonic commodity can be combined with the aggregate prices and quantities of nonhedonic commodities using normal index number theory.

We conclude this section by discussing one other aspect of the Silver and Heravi paper: namely, their use of matched model superlative indexes. A matched model price index for the hedonic commodity between periods t and $t + 1$ is constructed as follows. Let $I(t, t + 1)$ be the set of models k that are sold in both periods t and $t + 1$. Then the matched model Laspeyres and Paasche price indexes going from period t to period $t + 1$, P_L and P_P respectively, are

$$(57) \qquad P_L^t \equiv \frac{\sum_{k \in I(t,t+1)} P_k^{t+1} q_k^t}{\sum_{k \in I(t,t+1)} P_t^k q_t^k} ;$$

$$(58) \qquad P_P^t \equiv \frac{\sum_{k \in I(t,t+1)} P_k^{t+1} q_k^{t+1}}{\sum_{k \in I(t,t+1)} P_k^t q_k^{t+1}} .$$

In the above matched model indexes, we compare only models that were sold in both periods under consideration. Thus we are throwing away some of our price information (on prices that were present in only one of the two periods). The matched model superlative Fisher Ideal price index going from period t to $t + 1$ is $P_F^t \equiv (P_L^t P_P^t)^{1/2}$; that is, it is the square root of the product of the matched model Laspeyres and Paasche indexes. Now it is possible to compare the matched model Fisher measure of price change going from period t to $t + 1$, P_F^t, to the corresponding measure of aggregate price change that we could get from our hedonic model, which is ρ_{t+1}/ρ_t. We would hope that these measures of price change would be quite similar, particularly if the proportion of matched models is high for each period (as it is for the Silver and Heravi data). Silver and Heravi make this comparison for their hedonic models and find that the matched Fisher ends up about 2 percent lower for their U.K. washing machine data for 1998 compared to the hedonic models. It seems quite possible that this relatively large discrepancy could be due to the fact that the Silver and Heravi hedonic func-

---

44. If we have data for the q_k^t, then it is best to run sales-weighted regressions, as was discussed in the previous section. If we do not have complete market data on individual model sales but we do have total sales in each period, then we can run the hedonic regression model in equation (14) using a sample of model prices and then divide period t sales by our estimated ρ_t parameter in order to obtain an estimator for Q_t.

tional forms are only capable of providing a first-order approximation to arbitrary hedonic preferences, whereas the superlative indexes can provide a second-order approximation, and thus substitution effects are bigger for the superlative matched model price indexes.[45]

Thus an important implication of the Silver and Heravi paper emerges: it is not necessary to undertake a hedonic study if the following conditions hold:

- Detailed data on the price and quantity sold of each model are available.
- Between consecutive periods, the number of new and disappearing models is small, so that matching is relatively large.

We turn now to our final topic: a discussion of the additional problems that occur if we relax the assumption that the hedonic subutility function $f(z)$ is time invariant.

10.7 Changing Tastes and the Hedonic Utility Function

Several economists have suggested that there are good reasons why the hedonic utility function $f(z)$ introduced in section 10.2 may depend on time t.[46] In this section, we consider what changes need to be made to our basic hedonic model outlined in section 10.2 if we replace our time invariant hedonic utility function $f(\mathbf{z})$ by one that depends on time, say $f^t(\mathbf{z})$.[47]

If we replace our old $f(\mathbf{z})$ in section 10.2 with $f^t(\mathbf{z})$ and make the same other assumptions as we made there, we find that instead of our old equation (14), we now end up with the following equations.

$$(59) \qquad P_k^t = \rho_t f^t(\mathbf{z_k^t}); \quad t = 1, \ldots, T; k = 1, \ldots, K^t.$$

Up to this point, nothing much has changed from our previous 10.2 model that assumed a time-invariant hedonic subutility function $f(\mathbf{z})$, except that our new subutility function $f^t(\mathbf{z})$ will naturally have some time-

45. In favor of this interpretation is the fact that the matched model Laspeyres index was roughly the same as the hedonic indexes computed by Silver and Heravi. However, there are other factors at work, and this "explanation" may well be incomplete.

46. More precisely, Silver (1999a) and Pakes (2001) make very strong arguments (based on industrial organization theory) that the hedonic regression coefficients that are estimated using period t data should depend on t. Griliches (1961) also argued that the hedonic regression coefficients were unlikely to be constant over periods.

47. Before we proceed to our general discussion of time-dependent hedonic aggregator functions $f^t(\mathbf{z})$, we note a simple method originally due to Court (1939) and Griliches (1961) for allowing for time dependence that does not require any new methodology: simply use the previous time-independent methodology, but restrict the regression to two consecutive periods. This will give us a measure of overall price change for the hedonic commodity going from period t to $t + 1$, say. Then run another hedonic regression using only the data for periods $t + 1$ and $t + 2$, which will give us a measure of price change going from period $t + 1$ to $t + 2$. And so on.

dependent parameters in it. However, there is another major change that is associated with our new model, equation (59). Recall that in the time-invariant models discussed in section 10.3, we required only *one* normalization on the parameters, like $\rho_1 = 1$. In our new time-dependent framework, we require a normalization on the parameters in equation (59) for each period; that is, we now require T normalizations on the parameters instead of one in order to identify the ρ_t and the α parameters that characterize $f^t(\mathbf{z})$.

The simplest way to obtain the required normalizations is to make the hypothesis that the utility that *a reference model* with characteristics $\mathbf{z}^* \equiv (z_1^*, \ldots, z_N^*)$ gives the consumer the *same utility* across all periods in the sample. If we choose this reference utility level to be unity, then this hypothesis translates into the following restrictions on the parameters of $f^t(\mathbf{z})$:

$$(60) \qquad\qquad f^t(\mathbf{z}^*) = 1; \quad t = 1, \ldots, T.$$

Equations (59) and (60) now become our basic system of hedonic regression equations and replace our old system, equation (14) plus the normalization $\rho_1 = 1$.[48]

How should we choose the functional form for $f^t(\mathbf{z})$? Obviously, there are many possibilities. However, the simplest possibility (and it is the one chosen by Silver and Heravi) is to allow the α_n parameters that we defined for various functional forms in section 10.3 to depend on t; that is, the α_n defined in section 10.3 is replaced by α_n^t, and each period t parameter set is estimated by a hedonic regression that uses *only* the price and characteristics data for period t.[49] We leave to the reader the details involved in reworking our old algebra in section 10.3, changing the α_n into α_n^t and imposing the normalizations in equation (60) in place of our old normalization, $\rho_1 = 1$.

So far, so good. It seems that we have greatly generalized our old "static" hedonic model at virtually no cost. However, there is a hidden cost. Our new system of regression equations, (59) and (60), is in general *not invariant to the choice of the reference model with characteristics vector* $\mathbf{z}^*$. Thus if we choose a different reference model with characteristics vector $\mathbf{z}^{**} \neq \mathbf{z}^*$ and replace the normalizations in equation (60) with

48. If we define the imputed price of the reference model in period t as P^{t*}, it can be seen using equations (60) and (61) that $P^{t*} = \rho_t$ for $t = 1, \ldots, T$. Now in actual practice, when unrestricted period t hedonic regressions are run in isolation, researchers omit the time dummy and just regress, say, $\ln P_k^t$ on $\ln f^t(z_k^t)$, where the right-hand-side regression variables have a constant term. Then the researcher estimates the period t aggregate price of the hedonic commodity as $\rho^{t*} \equiv f^t(\mathbf{z}^*)$ where $\mathbf{z}^*$ is a conveniently chosen vector of reference characteristics. This procedure is equivalent to our time dummy procedure using the normalizations shown in equation (61).

49. If quantity sales data are available, then we recommend the weighted regression approach explained in section 10.5; recall equations (55). Also, in this case, if models are sold at more than one price in any given period, then we could weight each distinct price by its sales at that price or simply aggregate over sales of the specific model k in period t and let P_k^t be the unit value price over all of these sales. In what follows, we assume that the second alternative is chosen.

(61) $$f^t(\mathbf{z}^{**}) = 1; \quad t = 1, \ldots, T.$$

then in general, the new estimates for the aggregate hedonic commodity prices ρ_t will change. Thus the cost of assuming a time-dependent hedonic utility function is a lack of invariance in the relative prices of the aggregate hedonic commodity over time to our utility function normalization equations (60) or (61).

This lack of invariance in our estimated ρ_t need not be a problem for statistical agencies, provided that we can agree on a "reasonable" choice for the reference model that is characterized by the characteristics vector $\mathbf{z}^*$, since the important factor for the agency is to obtain "reasonable" and reproducible estimates for the aggregate hedonic commodity prices. Based on some discussion of this problem in Silver (1999b, 47), a preliminary suggestion is that we take $\mathbf{z}^*$ to be the sales-weighted average vector of characteristics of models that appeared during the sample period:

(62) $$\mathbf{z}^* \equiv \frac{\sum_{k=1}^{K} \sum_{t=1}^{T} q_k^t \mathbf{z_k}}{\sum_{k=1}^{K} \sum_{t=1}^{T} q_k^t},$$

where we have reverted to the notation used in section 10.6; that is, K is the total number of distinct models that we sold in the market over all T periods in our sample, and q_k^t is the number of models that have the vector of characteristics $\mathbf{z_k}$ that were sold in period t.[50]

Thus, once we pick functional forms for the $f^t(\mathbf{z})$ and add stochastic terms to equation (59), equations (60) and (61) and definition (62) completely specify our new hedonic regression framework. Of course, we still recommend that quantity weights (if available) be used in the econometric estimation for reasons explained in section 10.5; recall equation (54).

However, if the number of time periods in our sample T is large, then there is a danger that the overall characteristics vector $\mathbf{z}^*$ defined by equation (62) may not be very representative for any one or two consecutive periods. Thus we now suggest a different method of normalizing or making comparable the time dependent hedonic utility functions $f^t(\mathbf{z})$ that will deal with this lack of representativity problem. For each time period t, define $\mathbf{z^{t*}}$ to be the sales-weighted average vector of characteristics of models that appeared during period t:

(63) $$\mathbf{z^{t*}} \equiv \frac{\sum_{k=1}^{K} q_k^t \mathbf{z_k}}{\sum_{k=1}^{K} q_k^t}; \quad t = 1, \ldots, T.$$

Recall our basic hedonic regression equation (59), $P_k^t = \rho_t f^t(\mathbf{z_k^t})$. Now make the following normalizations:

50. If quantity information on sales of models, q_k^t, is not available, then define $\mathbf{z}^*$ as an unweighted arithmetic mean of the z_k.

(64) $$\rho_t = 1; \quad t = 1, \ldots, T.$$

Assuming that the parameters of the period t hedonic utility functions $f^t(\mathbf{z})$ have been estimated, we can now define the period t to $t + 1$ Laspeyres-, Paasche-,[51] and Fisher-type hedonic price indexes, respectively, as follows:

(65) $$P_L^{t,t+1} \equiv \frac{f^{t+1}(\mathbf{z}^{t*})}{f^t(\mathbf{z}^{t*})}; \quad t = 1, \ldots, T-1;$$

(66) $$P_P^{t,t+1} \equiv \frac{f^{t+1}(\mathbf{z}^{t+1*})}{f^t(\mathbf{z}^{t+1*})}; \quad t = 1, \ldots, T-1;$$

(67) $$P_F^{t,t+1} \equiv (P_L^{t,t+1} P_P^{t,t+1})^{1/2}; \quad t = 1, \ldots, T-1.$$

The Fisher-type hedonic price index is our preferred index. It can be seen that the Laspeyres and Paasche indexes defined by equations (65) and (66) can be quite closely related to Feenstra's upper and lower bounding indexes to his true index (and this superlative exact hedonic methodology is used by Silver and Heravi), depending on what functional form for f^t is chosen.

Once the parameters that characterize the time-dependent hedonic utility functions $f^t(\mathbf{z})$ have been estimated along with the associated aggregate period t hedonic commodity prices ρ_t,[52] then we can define period t aggregate demand for the hedonic commodity by[53]

(68) $$Q_t \equiv \sum_{k=1}^{K} f^t(\mathbf{z_k})q_k^t; \quad t = 1, \ldots, T.$$

The above model is our suggested *direct method* for forming exact aggregate period t prices and quantities, ρ_t and Q_t, for the hedonic commodity.

It is possible to use the outputs of hedonic regressions in another, more indirect way, along with normal index number theory, in order to construct aggregate price and quantity indexes for the hedonic commodity.[54] Recall equations (57) and (58) in the previous section, which defined the matched model Laspeyres and Paasche price indexes over hedonic models going from period t to $t + 1$. The problem with these indexes is that they throw

51. Berndt, Griliches, and Rappaport (1995, 262–63) and Berndt and Rappaport (2001) define the Laspeyres- and Paasche-type hedonic indexes in this way. However, the basic idea dates back to Griliches (1971, 59) and Dhrymes (1971, 111–12). Note that equations (66) and (67) break down if the vector of characteristics in period t is totally different from the vector of characteristics in period $t + 1$. Similarly, problems can arise if some characteristics are zero in one period and nonzero in another period; recall the log of zero problem discussed in section 10.3 above.

52. In our second method, in which we set the ρ_t equal to unity, define $\rho_1 = 1$ and $\rho_t + 1 = \rho_t P_F^{t,t+1}$ or $t = 1, 2, \ldots, T-1$ where the Fisher-type hedonic chain index $P_F^{t,t+1}$ is defined by equation (68). In this second method, once the aggregate prices ρ_t have been determined, we obtain the aggregate quantities Q_t as the deflated values, $\sum_{k=1}^{K} P_k^t q_k^t / \rho_t$, rather than using equations (69).

53. If quantity weights are not available, then we cannot compute Q_t.

54. See Moulton (1996, 170) for an exposition of these methods.

away information on models that are sold in only one of the two periods under consideration. One way of using this discarded information is to use the hedonic regressions in order to *impute* the missing prices.[55]

Suppose that model k was either unavailable or not sold in period t (so that $q_k^t = 0$) but that it was sold during period $t + 1$ (so that P_k^{t+1} and q_k^{t+1} are positive). The problem is that we have no price P_k^t for this model in period t, when it was not sold. However, for period $t + 1$, our hedonic regression equation for this model is the following equation (neglecting the error term):

$$(69) \qquad P_k^{t+1} = \rho_{t+1} f^{t+1}(\mathbf{z_k}).$$

Now we can use the estimated period $t + 1$ hedonic utility function f^{t+1} and the estimated period t aggregate price for the hedonic commodity, ρ_t, in order to define an imputed price for model k in period t as follows:

$$(70) \qquad P_k^{t*} \equiv \rho_t f^{t+1}(\mathbf{z_k}) = \rho_t \left(\frac{P_k^{t+1}}{\rho_{t+1}} \right) \text{ using (69)} = \left(\frac{\rho_t}{\rho_{t+1}} \right) P_k^{t+1}.$$

Thus the imputed price for model k in period t, P_k^{t*}, is equal to the observed model k price in period $t + 1$, P_k^{t+1}, times the reciprocal of the estimated rate of overall change in the price of the hedonic commodity going from period t to $t + 1$, $(\rho_t/\rho_t + 1)$.

Now suppose that model k sold in period t (so that P_k^t and q_k^t are positive) but that model k either disappeared or was not sold in period $t + 1$ (so that P_k^{t+1} is 0). The problem is that we have no price P_k^{t+1} for this model in period $t + 1$, when it was not sold. However, for period t, our hedonic regression equation for model k is the following equation (neglecting the error term):

$$(71) \qquad P_k^t = \rho_t f^t(\mathbf{z_k}).$$

Now we can use the estimated period t hedonic utility function f^t and the estimated period $t + 1$ aggregate price for the hedonic commodity, ρ_{t+1}, in order to define an imputed price for model k in period $t + 1$ as follows:

$$(72) \qquad P_k^{t+1*} \equiv \rho_{t+1} f^t(\mathbf{z_k}) \quad = \rho_{t+1} \left(\frac{P_k^t}{\rho_t} \right) \text{ using (71)} = \left(\frac{\rho_{t+1}}{\rho_t} \right) P_k^t.$$

Thus the imputed price for model k in period $t + 1$, P_k^{t+1*}, is equal to the observed model k price in period t, P_k^t, times the estimated rate of overall change in the price of the hedonic commodity going from period t to $t + 1$, (ρ_{t+1}/ρ_t).[56]

Now we can use the imputed prices defined by equations (70) and (72) in order to obtain price and quantity information on *all* models that were pres-

---

55. See Armknecht and Maitland-Smith (1999) for a nice review of imputation methods.

56. I believe that the approach outlined here is consistent with the approach used by Silver and Heravi to generate imputed prices for missing models. Triplett (2001) outlines other approaches.

ent in one or both of periods t and $t + 1$ and hence we can calculate the following completely matched Laspeyres and Paasche price indexes:

$$(73) \qquad P_L^t \equiv \frac{\sum_{k=1}^K P_k^{t+1} q_k^t}{\sum_{k=1}^K P_k^t q_k^t};$$

$$(74) \qquad P_P^t \equiv \frac{\sum_{k=1}^K P_k^{t+1} q_k^{t+1}}{\sum_{k=1}^K P_k^t q_k^{t+1}}$$

where we use the imputed price P_k^{t*} defined by equation (70) in place of the missing P_k^t if $q_k^t = 0$ but q_k^{t+1} is positive and we use the imputed price P_k^{t+1*} defined by equation (72) in place of the missing $P_k^{t+'}$ if $q_k^{t+1} = 0$ but q_k^t is positive.[57] Comparing our new Laspeyres and Paasche price indexes defined by equation (73) and (74) to our old matched model Laspeyres and Paasche price indexes defined by equations (57) and (58), it can be seen that our new indexes do not throw away any relevant price and quantity information and hence can be expected to be more "accurate" in some sense.

10.8   Conclusion

A number of tentative conclusions can be drawn from the Silver and Heravi (2001) paper and this discussion of it:

- Traditional superlative index number techniques that aggregate up model data based on matched models can give more or less the same answer as a hedonic approach, provided that the amount of matching is relatively large.
- Linear hedonic regressions are difficult to justify on theoretical grounds (at least based on our highly simplified approach to hedonic regressions) and hence should be avoided if possible.
- If completely unconstrained hedonic regressions are run on the data of each period, then care should be taken in the choice of a reference model that allows us to compare the utility of the hedonic commodity across periods. In particular, the estimates of aggregate price change in the hedonic commodity will in general not be invariant to the choice of the reference model.
- The use of quantity weights in hedonic regression models is strongly recommended if possible.
- Under certain conditions, if models are matched in each period, then the hedonic regression approach will give exactly the same answer as a

57. Obviously, if both q_k^t and q_k^{t+1} are zero, then we do not require estimators for the missing prices P_k^t and P_k^{t+1} in order to compute the Laspeyres and Paasche indexes defined by equations (74) and (75).

traditional statistical agency approach to the calculation of an elementary index.
- We have not achieved a consensus on exactly what the "best practice" hedonic regression specification should be, but flexible functional form considerations should probably be a factor in the discussion of this problem.

References

Armknecht, P. A., and F. Maitland-Smith. 1999. Price imputation and other techniques for dealing with missing observations, seasonality, and quality change in price indices. In *Proceedings of the measurement of inflation conference,* ed. M. Silver and D. Fenwick, 25–49. London: Office for National Statistics.

Berndt, E. R. 1991. *The practice of econometrics: Classic and contemporary.* Reading, Mass.: Addison-Wesley.

Berndt, E. R., Z. Griliches, and N. J. Rappaport. 1995. Econometric estimates of price indexes for personal computers in the 1990's. *Journal of Econometrics* 68:243–68.

Berndt, E. R., and N. J. Rappaport. 2001. Price and quality of desktop and mobile personal computers: A quarter century historical overview. *The American Economic Review* 91 (2): 268–73.

Christensen, L. R., D. W. Jorgenson, and L. J. Lau. 1975. Transcendental logarithmic utility functions. *American Economic Review* 65:367–83.

Court, A. T. 1939. Hedonic price indexes with automotive examples. In *The dynamics of automobile demand,* 99–117. New York: General Motors Corporation.

Davies, G. R. 1924. The problem of a standard index number formula. *Journal of the American Statistical Association* 19:180–88.

Dhrymes, P. J. 1971. Price and quality changes in consumer capital goods: An empirical study. In *Price indexes and quality change,* ed. Z. Griliches, 88–149. Cambridge, Mass.: Harvard University Press.

Diewert, W. E. 1971. An application of the shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy* 79:481–507.

———. 1974. Applications of duality theory. In *Frontiers of quantitative economics,* vol. 2, ed. M. D. Intriligator and D. A. Kendrick, 106–71. Amsterdam: North-Holland.

———. 1993. Duality approaches to microeconomic theory. In *Essays in index number theory,* vol. 1, ed W. E. Diewert and A. O. Nakamura, 105–75. Amsterdam: North-Holland.

———. 1995. Axiomatic and economic approaches to elementary price indexes. Discussion Paper 95-01. Vancouver, Canada: University of British Columbia, Department of Economics. Available at [http://web.arts.ubc.ca/econ/diewert/hmpgdie.htm].

———. 1998. Index number issues in the Consumer Price Index. *The Journal of Economic Perspectives* 12:47–58.

Feenstra, R. C. 1995. Exact hedonic price indices. *Review of Economics and Statistics* 77:634–54.

Fixler, D., and K. D. Zieschang. 1992. Incorporating ancillary measures of process

and quality change into a superlative productivity index. *The Journal of Productivity Analysis* 2:245–67.

Goldberger, A. A. 1968. The interpretation and estimation of Cobb-Douglas functions. *Econometrica* 35:464–72.

Griliches, Z. 1961. Hedonic price indexes for automobiles: An econometric analysis of quality change. In *Price indexes and quality change,* ed. Z. Griliches, 55–87. Cambridge, Mass.: Harvard University Press.

———. 1971. Introduction: Hedonic price indexes revisited. In *Price indexes and quality change,* ed. Z. Griliches, 3–15. Cambridge, Mass.: Harvard University Press.

Hicks, J. R. 1946. *Value and capital.* 2d ed. Oxford, U.K.: Clarendon Press.

Kokoski, M. F., B. R. Moulton, and K. D. Zieschang. 1999. Interarea price comparisons for heterogeneous goods and several levels of commodity aggregation. In *International and interarea comparisons of income, output, and prices,* Studies in Income and Wealth, vol. 61, ed. A. Heston and R. E. Lipsey, 123–66. Chicago: University of Chicago Press.

Koskimäki, T., and Y. Vartia. 2001. Beyond matched pairs and Griliches-type hedonic methods for controlling quality changes in CPI sub-indices. Paper presented at the sixth Ottawa Group Meeting. 2–6 April, Canberra, Australia.

Muellbauer, J. 1974. Household production theory, quality, and the "hedonic technique." *The American Economic Review* 64 (6): 977–94.

Moulton, B. 1996. Bias in the Consumer Price Index: What is the evidence? *Journal of Economic Perspectives* 10 (4): 139–77.

Pakes, A. 2001. Some notes on hedonic price indices, with an application to PCs. Paper presented at the NBER Productivity Program Meeting. 16 March, Cambridge, Massachusetts.

Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 82 (1): 34–55.

Scheffé, H. 1959. *The analysis of variance.* New York: John Wiley and Sons.

Silver, M. 1995. Elementary aggregates, micro-indices, and scanner data: Some issues in the compilation of consumer prices. *Review of Income and Wealth* 41:427–38.

———. 1999a. Bias in the compilation of consumer price indices when different models of an item coexist. In *Proceedings of the fourth meeting of the International Working Group on Price Indices,* ed. W. Lane, 21–37. Washington, D.C.: Bureau of Labor Statistics.

———. 1999b. An evaluation of the use of hedonic regressions for basic components of consumer price indices. *The Review of Income and Wealth* 45 (1): 41–56.

Triplett, J. 2001. *Handbook on quality adjustment of price indexes for information and communication technology products.* Paris: Organization for Economic Cooperation and Development. Forthcoming.

Walsh, C. M. 1901. *The measurement of general exchange value.* New York: Macmillan and Company.

———. 1921. *The problem of estimation.* London: P. S. King and Son.

Yu, K. 2001. Trends in Internet access prices in Canada. Paper presented at the sixth Ottawa Group Meeting. 2–6 April, Canberra, Australia.