Overcoming Data Limitations in Hedonic Regressions

by

Ludwig v. Auer*
Otto-von-Guericke-Universität Magdeburg, Germany

(October 2004)

Abstract:

Recently, the use of hedonic regressions have become commonplace both inside and outside of national statistical agencies. This increased popularity is due in part to the availability of scanner data. Scanner data is often costly, however, and for some purposes does not even exist. In these instances, other data sources must be found (e.g., catalogues, magazine test results, etc.). Unfortunately, these alternative sources often provide small sample sizes and the hedonic regressions conducted require extremely restrictive assumptions to be placed on the estimated coefficients. This paper evaluates these assumptions and presents an alternative procedure that exhibits several appealing features. The robustness of this alternative method is examined.

Keywords: Hedonic Regression, Price Index, Measurement Bias, Technical Progress;
JEL Classification: C23, C43, L63.

* I am very grateful to Daniel Melser for his valuable comments on a related paper and to Kim Zieschang for his stimulating critique. I am particularly indebted to John E. Brennan whose advice greatly improved the readability of this study. All errors and omissions, of course, are my own.

1. Introduction

In technologically advanced economies, goods and services are undergoing a continuous improvement in their quality and performance characteristics. This poses a great problem for the accurate measurement of price-trends and thus the assessment of overall economic performance in these countries. The impact that quality change can have on price measurement is a problem that has been recognized for some time. In his famous study, Laspeyres (1871) questioned the reliability of the reported import prices (*Hamburger Börsen-Preiscourant*) that he utilized for the computation of his price index. He wrote, "*... we also do not know ... whether the increased price 1856-60 per centner is due to the rise in price of the same quality or due to the importation of a better quality*" (p. 306, the translation is provided by the author).

Today, widespread agreement exists that quality change accounts for a major part of the potential bias in the measurement of prices. For some time, many statistical agencies have computed their quality-adjusted prices utilizing the so-called "matched model method". Various studies based on scanner data, however, have shown that the matched model method may not provide an accurate measure of quality change (e.g., Silver and Heravi, 2004). For this reason, the hedonic regression approach has been advocated as an alternative to the matched model method. Over the past decade, therefore, the use of hedonic regressions has become quite common both inside as well as outside of the national statistical agencies.

Part of the popularity that hedonic regressions have enjoyed is certainly due to the improved access to scanner data. Scanner data, however, are often costly to obtain and for some purposes they do not even exist. For use in these studies, other data sources must be utilized (e.g., product catalogues, magazine test results, etc.). Unfortunately, these data sources often only offer very limited sample sizes. As a result, the associated hedonic regressions usually require restrictive assumptions to be placed on the coefficients to be estimated. This paper evaluates the assumptions typically made and describes an alternative approach that exhibits several appealing features.

This paper proceeds as follows. In Section 2, the hedonic regression procedure is described that can be utilized when sufficient data are available. Section 3 outlines the two traditional hedonic procedures that are widely used when data shortage problems exist. In Section 4, an alternative method is described that avoids some of the shortcomings of the two traditional hedonic procedures. Section 5 presents the results of an empirical application of this new technique and it examines the robustness of these results. Concluding remarks are contained in Section 6.

2.Hedonic Regression with Sufficient Data

The products that are typically studied using hedonic regression techniques are products from the Information Technology (IT) sector or consumer durables such as washing machines and television sets. In the market place many different models of these products are being sold. Consider, for example, some specific product (e.g., automobile tires). In hedonic regressions the price of model *i* during time period *t*, p_i^t, is related to *K* of its qualitative characteristics, q_{ki} $(k = 1,2,\ldots,K)$. The most commonly used functional relationship is the semi-logarithmic form:

$$\ln p_i^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t q_{ki} + u_i^t \tag{1}$$

The dependent variable in this equation, $\ln p_i^t$, is the natural logarithm of the observed price of model *i* during time period *t* $(t = 1,2,\ldots,T)$. The qualitative characteristics of model *i* are represented by the variables q_{ki} $(k = 1,2,\ldots,K)$. The stochastic variable u_i^t is a random error term that is assumed to conform to all of the assumptions commonly made in regression analysis. The intercept term, β_0^t, can be interpreted as measuring those factors in the marketplace that determine the logarithmic price of model *i* in time period *t* other than the model's own inherent product quality features. The slope coefficients, β_k^t $(k = 1,2,\ldots,K)$, on the other hand, can be interpreted as the semi-elasticities of a model's price with respect to the k^{th} qualitative characteristic.

Usually, in any arbitrarily selected time period *t*, several different models will be available for sale. During the following time period $(t + 1)$, however, only a portion of these models might still be available. On the other hand, in time period $(t + 1)$ some new models will become available that were not yet available in time period *t*. A hypothetical market evolution describing twelve monthly time periods is illustrated in Figure 1. The Figure shows $N = 149$ observations.

Here Figure 1

When scanner data is available, one usually has detailed information concerning the qualitative characteristics of several hundred different models being sold in the market place. Furthermore, for each time period *t* and for all models *i*, information is available detailing the quantities sold and the associated transaction prices. This allows the researcher to separately estimate the T hedonic equations specified in expression (1). In ILO *et al.* (2004), it is argued that this should be done on the basis of weighted least squares, where the weights assigned to specific models should reflect the expenditure shares associated with these models.

Whatever estimation technique is applied, estimated values of $\hat{\beta}_0^t$ and $\hat{\beta}_k^t$ ($k = 1,2,\ldots,K$) will be generated. Based on these estimated values one can calculate for each period t and each model j the predicted prices $\hat{p}_j^t$, regardless of whether model j was available in period t or not. The average price change between two periods t and (t + 1) can be computed by averaging the price ratios $\hat{p}_j^{t+1} / \hat{p}_j^t$ of those models j that were available in at least one of the two periods. For this purpose, a weighted geometric mean appears to be the preferable procedure where the weights reflect the expenditure shares of model j.

The heretofore-outlined procedure, however, requires a large amount of available data. For some products, scanner data can provide such a database. However, suitable scanner data are available only for a subset of the products undergoing technical changes. For example, for services hardly any scanner data exist. Moreover, it is difficult to obtain scanner data that reaches back as far as ten or twenty years. Therefore, the calculation of long-term price-trends usually must be conducted without the use of scanner data. Even if suitable scanner data did exist, it is often too costly to implement them into a routine price index calculation procedure. These problems suggest that there is a need to search for alternative hedonic estimation procedures that can be utilized in the context of small sample sizes when scanner data or other large databases are unavailable. Such procedures are discussed in the following section.

3. Hedonic Regression with Insufficient Data: TDV and AYR Method

Figure 1 illustrates the hypothetical market evolution for some specific product. If scanner data is unavailable and a database has to be constructed from some other source (e.g., published catalogues, consumer magazine test results, etc.), then possibly data will be available for only a limited number of the models being sold in the marketplace. The available database may appear like the one depicted in Figure 2. In this figure, the black squares, ■, signify the recorded data contained in the database. The empty squares, □, are the models that are not recorded in the database. Only N = 28 of the possible 149 observations are recorded in this case.

Here Figure 2

For some of the months no observations are available and for others only a few observations exist. Suppose, for example, that the number of qualitative characteristics is K = 3. In this case, Equation (1) could be estimated only in three of the monthly periods and the results would unlikely be robust. The simplest “solution” to this data shortage problem is to assume that all of the slope coefficients, β_k^t (k = *1,2,…,K*), are constant over all T periods. This is the so-called Time Dummy Variable (TDV) method. Based on this assumption, the T relationships expressed in Equation (1) can be pooled into a single equation for estimation:

$$\ln p_i = \beta_0 + \sum_{t=2}^{T} \gamma^t d_i^t + \sum_{k=1}^{K} \beta_k q_{ki} + u_i \tag{2}$$

with

$d_i^t = 1$ if model i is observed in time period t

$d_i^t = 0$ if model i is not observed in time period t.

Utilizing scanner data, Grient (2004) reported that the slope coefficients usually change within a year. If the T pooled time-periods are monthly, then the TDV method’s simple pooling procedure is based upon invalid assumptions. As a consequence, the TDV method in its simple form is questionable.

There exists, however, an alternative procedure that utilizes the concept of the TDV method in a more flexible way. This procedure is known as the Adjacent Year Regression (AYR) method. The basic concept behind the AYR method is to pool the observations from two or more consecutive time periods (e.g., months or years) into a sub-sample and to assume that the slope coefficients remain constant throughout these pooled time-periods. Suppose, for example, that the observations relating to time periods t = 1 through t = 4 are pooled into a sub-sample. This sub-sample is then estimated using Equation (2) with T = 4. Next, the observations from the time periods t = 4 through t = 7 are pooled into a second sub-sample such that there exists an overlap between the first and the second sub-sample where t = 4. This second sub sample is also estimated using Equation (2).

The estimated overall percentage price changes between periods 1, 2, 3, and 4 are obtained from the estimated coefficients $\hat{\gamma}^2$, $\hat{\gamma}^3$, and $\hat{\gamma}^4$. The numerical values of these estimated coefficients are obtained from the hedonic regression conducted on the first sub-sample. The estimated overall percentage price changes between periods 4, 5, 6, and 7 are equal to $\hat{\gamma}^5$, $\hat{\gamma}^6$, and $\hat{\gamma}^7$. These estimated values are obtained from the regression on the second sub-sample. Therefore, the overlapping pooled regression procedure generates a time series of estimated coefficients $\hat{\gamma}^2$, $\hat{\gamma}^3$, … , $\hat{\gamma}^7$.

Following the same procedure, overlapping pooled regressions can also be conducted for any subsequent sub-samples. Finally, a time series of estimated coefficients $\hat{\gamma}^2$, $\hat{\gamma}^3$, … , $\hat{\gamma}^T$ is obtained. These estimates can be used to calculate the overall price-trend of the product:

$$I_t = 100 \cdot \prod_{s=1}^{t} \exp(\hat{\gamma}^s), \qquad t = 1,2, \dots ,T \tag{3}$$

where, by definition, $\hat{\gamma}^1 = 0$.

Unfortunately, however, the AYR method exhibits some theoretical shortcomings and it does not utilize all of the available information. One of the problems with the AYR method is caused by the overlapping nature of the sub-samples used in estimation. As is demonstrated in Auer and Brennan (2004), these overlapping time periods usually lead to estimation bias. In those exceptional cases where the AYR method does not produce biased estimated coefficients, the procedure is not efficient. Another problem is the fact that the AYR method often still suffers from too few observations. As a result, many of the estimated coefficients obtained are not robust. For an empirical example see Auer (2004).

The AYR method neglects some of the most valuable information contained within the database. Figure 2 illustrates this deficiency. Consider a sub-sample formed by pooling the time periods $t = 1$ through $t = 4$. Two of the models appear twice in the pooled sub-sample. A price change of these models can be purely attributed to the passage of time. This is precisely the type of information that one is looking for in quality-adjusted price-trend analysis. In the AYR method no coefficient exists that takes account of prices changes within the sub-sample. These price differences are treated as a disturbance that provides no information concerning the price change over time. The AYR method also neglects information that can be found outside of the database. Studies such as Grient (2004), Heravi and Silver (2004), and Mulligan (2003) utilize scanner data to show that the numerical values of the slope coefficients are not erratic, but rather change in a gradual manner over time. This important finding is not incorporated in the AYR method.

A more detailed discussion of these weaknesses can be found in Auer (2004). That study also introduces an alternative estimation procedure that avoids the deficiencies of the AYR method. This procedure is called the Continuously Changing Coefficient method (CCC) method and is presented in the following section.

4. Hedonic Regression with Insufficient Data: CCC Method

If it is assumed that Equation (1) represents a valid relationship for each time period t, then, as a consequence, a time series of values ($\beta_k^1, \beta_k^2, \ldots, \beta_k^T$) for each of the $K + 1$ coefficients would have to be estimated. The CCC method acknowledges that the numerical values of these coefficients can change over time. It is assumed, however, that these changes occur only gradually. Therefore, the coefficients in each of these K time series are modeled using a general polynomial form of degree Z_k:

$$\beta_k^t = (\beta_k + \sum_{z=1}^{Z_k} \theta_{kz} t^z), \qquad \text{with} \quad k = 0,1,2,...,\mathrm{K}, \tag{4}$$

where the parameters, θ_{kz}, are the coefficients of the polynomial and t represents a trend variable that runs from 1 to T. Polynomials of degree five, $Z_k = 5$, are capable of approximating very general time series patterns.

Utilizing the polynomials specified by Equation (4), the T relationships expressed in Equation (1) can be pooled into a single equation for estimation:

$$\ln p_i = \beta_0 + \sum_{z=1}^{Z_0} \theta_{0z} t^z + \sum_{k=1}^{K} (\beta_k + \sum_{z=1}^{Z_k} \theta_{kz} t^z) \, q_{ki} + u_i \; . \tag{5}$$

In this equation, the expression in brackets is a semi-elasticity that may change over time.

Estimation of Equation (5) generates the estimated coefficients $\hat{\beta}_k$ and $\hat{\theta}_{kz}$ (k = 0,1,..., K and z = 1,2,…, Z_k). These estimates can be used to calculate the predicted prices of each model j, $\hat{p}_j^t$, in each time period t. Similar to the case of hedonic regressions with sufficient data, these predicted prices $\hat{p}_j^t$ can be used to compute the overall price-trend of the product for the complete time span. For the transformation of predicted prices into overall price-trends various reasonable approaches are conceivable. Does it matter which of these approaches is used? If the predicted prices $\hat{p}_j^t$ were estimated using suitable scanner data, then the various approaches available would unlikely result in price-trends that deviate from each other in any significant way. However, when scanner data are not available and therefore the predicted prices $\hat{p}_j^t$ have to be estimated utilizing the CCC method, then the time trends associated with different approaches are more likely to differ from each other. Drawing on an empirical example, the following section explores several approaches and examines whether they lead to significant differences. It also examines whether in Equation (5) the degree of the polynomials has a significant impact on the price-trend computed from the predicted prices $\hat{p}_j^t$.

5. How Robust is the CCC Method?

Utilizing the CCC method, Auer (2004) estimated a quality-adjusted twelve-year price-trend for laser printers. The database used was derived from published test reports that appeared in a German computer magazine entitled *c't - magazin für computertechnik*. These reports contained information concerning the quality and performance characteristics of 232 different laser printer models sold under 20 different brand names. Some of the printer models considered were observed in two or three different months, increasing the number of observations to N = 378. These observations occurred during a time span of 144 months (January 1992 to December 2003). The structure of the database resembles that of Figure 2.

Eleven qualitative characteristics, q_{ki} (k = 1,2,...,11), and eight brand name dummy variables, b_{mi} (m = 1,2,...,8), were included in the estimation. The eight brand names included were those that had more than 14 observations in the complete database. The estimated hedonic regression can be written in the following form:

$$\ln \hat{p}_i = \hat{\beta}_0 + \sum_{z=1}^{Z_0} \hat{\theta}_{0z} t^z + \sum_{k=1}^{11} (\hat{\beta}_k + \sum_{z=1}^{Z_k} \hat{\theta}_{kz} t^z) q_{ki} + \sum_{m=1}^{8} (\hat{\delta}_m + \sum_{z=1}^{Z_m} \hat{\phi}_{mz} t^z) b_{mi} \; . \qquad (6)$$

In this equation $\hat{\theta}_{0z}$, $\hat{\theta}_{kz}$, and $\hat{\phi}_{mz}$ are the estimated coefficients of the polynomials. The other estimated coefficients are $\hat{\beta}_0$, $\hat{\beta}_k$, and $\hat{\delta}_m$. Numerical values for all of these estimated coefficients were provided by the least squares method. Using these estimated coefficients, an estimated price $\hat{p}_j^t$ was computed using Equation (6) for each of the 232 printer models j in each of the 144 monthly time periods considered. These calculations provide a time series of 144 estimated prices for each printer j. All of the 232 printers show different downward sloping price-trends.

The values of the estimated coefficients in each month, and therefore the estimated prices, are the most reliable for those printers that were actually sold in that month. Therefore, each of the 232 printers had a "time span of relevance" assigned to them. For example, the printer model NEC Superscript 660 was observed only once, namely in month t = 39 (March 1995). For this printer, the time span of relevance is defined to extend from month t = (39 – S) to month t = (39 + S), where S stands for "span". This time span is identical for all printers. Choosing S = 6 implies that the time span of relevance for the printer model NEC Superscript 660 extends from t = 33 (39 – 6), September 1994, to t = 45 (39 + 6), September 1995. The printer model Canon LBP 460 was observed twice, namely in month t = 49 and in month t =

55. Accordingly, this model's time span of relevance extends from month t = 43 (49 – 6), July 1995, to month t = 61 (55 + 6), January 1997.

Different laser printer models have different estimated price-trends. In order to obtain the average price change of laser printers between month t = 1 and month t = 2, some type of average of the individual price ratios $\hat{p}_j^2 / \hat{p}_j^1$ of those printer models with time spans of relevance covering both of these months must be computed. This average price ratio is denoted by $\hat{r}^2$. Subsequently, the heretofore-explained procedure can be accomplished for the months t = 2 and t = 3, and also for each subsequent pair of adjacent months. This procedure generates a time series of price ratios ($\hat{r}^2, \hat{r}^3, ..., \hat{r}^T$). The price ratio $\hat{r}^1$ is defined to be equal to one. The price trend of laser printers can be obtained from values of the price index,

$$I_t = 100 \cdot \prod_{s=1}^{t} \hat{r}^s , \qquad \text{with } t = 1,2, \dots ,T .$$

The procedure for estimating the overall price-trend leaves several areas of choice. Opting for one or the other alternative may affect the overall result. The rest of this study is devoted to this aspect. It examines three areas where decisions have to be made. More specifically, it examines the consequences arising from the choice

- of the polynomial order in Equation (6),
- of S (defining the time span of relevance), and
- of taking geometric or arithmetic averages of the individual price ratios $\hat{p}_j^{t+1} / \hat{p}_j^t$.

Figure 3 shows the price-trends associated with three different specifications of Equation (6). In all three specifications the polynomial order for the intercept coefficients is three (Z_0 = 3). However, the polynomial order for the other coefficients, Z_k (k = 1,2,...,11) and Z_m (m = 1,2,...,8), differs between the three specifications. The price trend denoted by Z = 1 results from applying a polynomial of order one ($Z_k = Z_m = 1$), the price trend Z = 2 is based on a polynomial of order two ($Z_k = Z_m = 2$), and the price trend Z = 3 results from a polynomial of order three ($Z_k = Z_m = 3$). The differences between these different price-trends are remarkably small. Interestingly, however, polynomials of order four or higher generated very implausible price-trends characterized by massive upward and downward trends. It must be left for future research to examine whether this finding also applies to other empirical applications.

Here Figure 3

The length of the time span of relevance for a printer model depends on the choice of S. The computations of the price-trends depicted in Figure 3 are all based on $S = 6$. How influential is the choice of S? In the laser printer example, the value of S affected the price-trend only very moderately. Figure 4 shows the price-trends associated with three different values of S, namely S = 2, 12, and 24.

Here Figure 4

The last of the three choices to be made is the averaging procedure for the price ratios $\hat{p}_j^{t+1} / \hat{p}_j^t$. Does an arithmetic average produce a price-trend that significantly deviates from a price-trend that is based on a geometric average? It turns out that the difference in each period is smaller than 0.4 percentage points. Therefore, the price trends are almost identical. It seems fair to conclude that the choice between a geometric and an arithmetic average is irrelevant – at least in the specific empirical application considered here.

All of these results suggest that in the present application the CCC method generates reasonably stable price trends. In view of the data limitations present in the laser printer database, this certainly comes as a positive surprise. The robustness of the CCC method becomes even more impressive, once one compares it to the robustness of the AYR method. As mentioned heretofore, some of the regressions performed using the AYR method suffered from too few observations. Auer (2004) reports that the inclusion or exclusion of a single observation can change the overall price-trend by as much as 10 percentage points. In the CCC method, however, the inclusion or exclusion of a single observation is hardly noticeable.

6. Concluding Remarks

This study discusses three possible procedures for performing hedonic regressions when data shortage problems exist. Two of them are well known techniques, namely the TDV and the AYR methods. The third one, the CCC method, has been introduced in the literature only recently. It is argued that the two traditional procedures, the TDV and AYR methods, are less suitable than is the CCC method. Applied to a database consisting of information concerning laser printers, the CCC method generates surprisingly robust price-trends. A subject for future research is to determine whether this result will prove true for other databases as well.

References

Auer, L. von (2004), Hedonic Quality Adjustment: The CCC Method, Otto-von-Guericke-Universität Magdeburg, FEMM Working Paper Series, Magdeburg.

Auer, L. von, J. E. Brennan (2004), Eliminating Estimation Bias in Quality Adjusted Hedonic Price Measurement, Otto-von-Guericke-Universität Magdeburg, FEMM Working Paper Series, Magdeburg.

Grient, H. van der (2004), Scanner Data on Durable Goods: Market Dynamics and Hedonic Time Dummy Price Indexes, Discussion Paper 04011, Voorburg / Heerlen: Statistics Netherlands.

Heravi, S., M. Silver (2004), On the Stability of Hedonic Coefficients and their Implications for Quality-Adjusted Price Change Measurement, published in these Conference Proceedings.

ILO, IMF, OECD, UNECE, Eurostat, The World Bank (2004), Consumer Price Index Manual: Theory and Practice, Geneva: International Labor Office.

Laspeyres, E. (1871), Die Berechnung einer mittleren Warenpreissteigerung, Jahrbücher für Nationalökonomie und Statistik, 296-314.

Mulligan, P. H. van (2003), Quality Aspects in Price Indices and International Comparisons: Applications of the Hedonic Method, PhD Thesis (University of Groningen), Statistics Netherlands, Voorburg.

Silver, M. and S. Heravi (2004), A Failure in the Measurement of Inflation: Results from a Hedonic and Matched Experiment Using Scanner Data, Journal of Business and Economic Statistics, forthcoming.

		model															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	1	■	■	■	■	■	■	■	■	■	■	■	■				
	2	■	■	■	■	■	■	■	■	■	■	■	■	■			
	3	■	■	■	■	■	■	■	■	■	■	■	■	■			
	4		■	■	■	■	■	■	■	■	■	■	■	■			
	5		■	■	■	■	■	■	■	■	■	■	■	■			
month	6			■	■	■	■	■	■	■	■	■	■	■	■		
	7			■	■	■	■	■	■	■	■	■	■	■	■		
	8			■	■	■	■	■	■	■	■	■	■	■	■		
	9			■	■	■	■	■	■	■	■	■	■	■	■	■	
	10				■	■	■	■	■	■	■	■	■	■	■	■	
	11				■	■	■	■	■	■	■	■	■	■	■	■	■
	12				■	■	■	■	■	■	■	■	■	■	■	■	■

Figure 1: The Available Models Within One Year.

		model															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	1	□	■	□	□	■	■	□	■	□	■	□	□				
	2	□	□	□	□	□	□	□	□	□	□	□	□	□			
	3	□	□	□	□	■	□	□	■	□	□	□	□	■			
	4		□	□	□	□	□	□	□	□	□	□	□	□			
	5		■	■	■	□	■	□	■	□	■	□	■	□			
month	6			□	□	□	□	□	□	□	□	□	□	□	□		
	7			□	□	■	□	□	□	□	□	□	□	■	□		
	8			□	□	□	□	□	□	□	□	□	□	□	□		
	9			□	■	□	□	□	□	□	□	■	□	□	■	□	
	10				□	□	□	□	□	□	□	□	□	□	□	□	
	11				■	□	■	□	■	□	■	□	■	■	■	■	□
	12				□	□	□	□	□	□	□	□	□	□	□	□	□

Figure 2: The Models Recorded Within One Year.

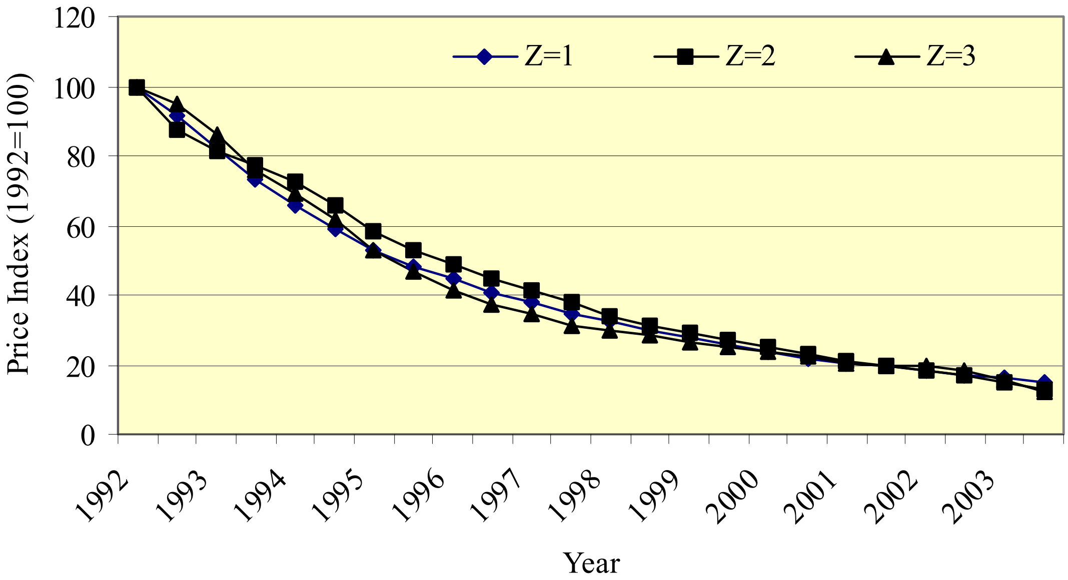

Figure 3: The Sensitivity to the Polynomial Order.

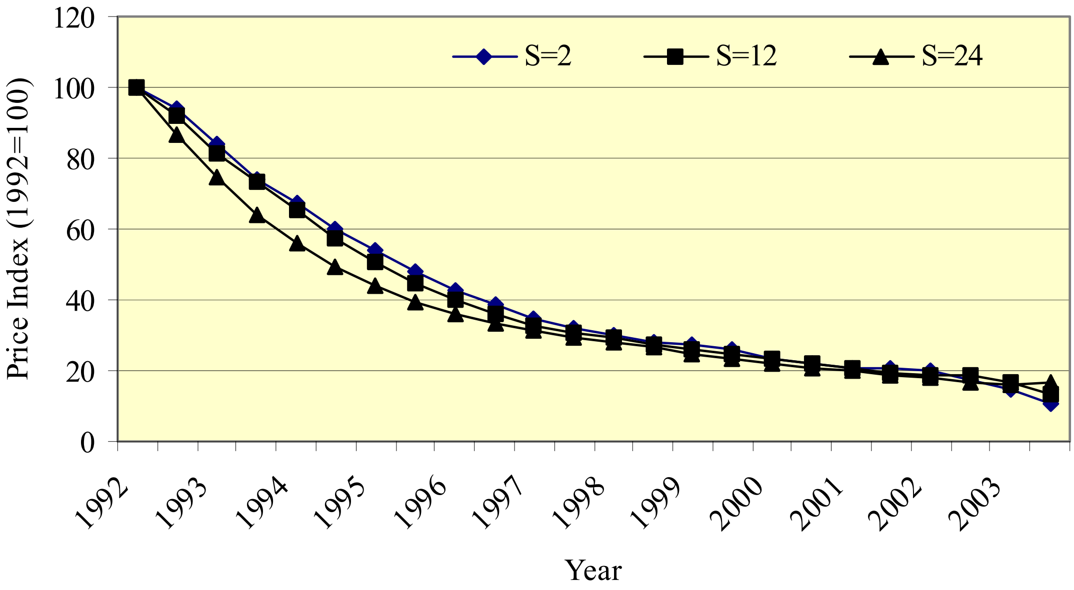

Figure 4: The Insensitivity to the “Time Span of Relevance”.