Quantity versus Quality of Characteristic Data in Hedonic Regressions

Teague Ruder*, Mick Silver and Ted To***

Draft June 1 2004

*Bureau of Labor Statistics
2 Massachusetts Ave., NE
Room 3105
Washington, DC 20212
USA

**Cardiff University
Colum Drive
Cardiff CF10 3EU
UK

Email: Ruder_T@bls.gov
To_T@bls.gov
Silver@cardiff.ac.uk

Abstract

A great deal of effort is expended by statistical agencies to estimate the "best" hedonic regression for the purpose of price adjustment. We argue here that resources used to attain the best specification may be better allocated elsewhere. Although coefficient estimates of hedonic regressions with omitted variables are biased, it is well-established that predicted prices are unbiased. Moreover, we show that quality adjusting an existing price using these biased coefficients will also be unbiased. However, these predicted prices and quality adjustments will be less precise than those using a more complete specification. Yet this loss in precision will be ameliorated to the extent that omitted characteristics are correlated with included characteristics. With fewer product characteristics, statistical agencies expend fewer resources in collecting and cleaning data and the quality of the characteristic data may be improved, further reducing the cost of data cleaning. We examine the effect of reducing the number of product characteristics using scanner data on washing machines from Great Britain.

I. Introduction

A great deal of effort is expended by statistical agencies to estimate the "best" hedonic regression for the purpose of price adjustment. We argue here that instead of concentrating so much effort on attaining the best specification, statistical agencies may better spend their resources by quality adjusting for fewer characteristics and focusing their efforts on improving price indices in other ways. Although the coefficient estimates of hedonic regressions with omitted variables are biased, it is well known that prices predicted using such hedonic regressions are unbiased. Moreover, we show that quality adjustments using these biased coefficients are also unbiased.[1]

Recognizing that quality adjustments using hedonic regressions with omitted characteristics are not biased, a number of benefits can be realized by estimating hedonic regressions with fewer characteristic variables. To begin with, data cleaning is an extremely time-consuming and costly activity. For the US Bureau of Labor Statistics (BLS) Economic Assistants (EA) are responsible for coding numerous characteristics for each product, amounting to, for example, y characteristics for washing machines. The BLS Commodity Analyst (CA) then identifies which individual models have undergone a change in quality, and applies a hedonic quality adjustment to the price. In order to make a quality adjustment to monthly price quotes, the CA looks up individual product models, compares and corrects product characteristics as coded by the EAs.

Reducing the list of characteristics to be collected reduces the cost of collecting price quotes. In addition, reducing the list of characteristics to be collected potentially improves the quality of the

[1] In the ensuing discussion we at times use the term "quality adjustment" to refer to both price prediction and explicit quality adjustment.

characteristic data collected. For example, at the BLS, the checklists used by EAs are complicated documents. Checklists have been designed to give EAs the flexibility to collect data on new characteristics, should they be observed. Unfortunately, this flexibility leads to some ambiguity and as a result EAs often differ in their interpretation of what quality attributes are to be collected. This causes inconsistencies in monthly price quotes that must be cleaned in a time-consuming and costly manner. With a shorter checklist, there will be less ambiguity regarding what characteristics need to be coded so that the quality of product characteristics may actually improve. With shorter checklists data collection is less costly *and* the data may be of higher quality, requiring less cleaning by CAs.

These savings come at the cost of reduced (statistical) efficiency of the estimates required for quality adjustments. However, this loss of efficiency can be ameliorated to some extent by relying on the multicollinearity of included and excluded characteristics. That is, the efficiency loss from dropping variables is reduced if there is multicollinearity between the dropped variables and the included variables. We would expect to find correlation in product characteristics. The literature on multiproduct monopoly price discrimination[2] can be reinterpreted as a model of characteristic bundling. This literature suggests that product characteristics will be sold in bundles and therefore one should expect them to be correlated. In other words, theory suggests that product characteristics will be correlated and as a result, the efficiency lost from dropping variables should not be as severe as one might expect.

To summarize the preceding discussion, there are substantial cost savings (both data collection and cleaning) to be realized by reducing the number of characteristic variables used for hedonic

[2] See for example, Adams and Yellen (1976), Armstrong (1996, 1999), McAfee *et al.* (1989), and Mirman and Sibley (1980).

price adjustment. These savings result in less precise price estimates from the hedonic regression and, in turn, quality adjustments to prices. But this loss in precision will be reduced if, as one would expect in theory, product characteristics are correlated.

II. Omitted variables and quality adjustment

In this section we first show that coefficients estimated using omitted variable specifications are biased and that the difference between predicted prices from an omitted variable specification and a full specification depends on the collinearity between the variables on the omitted and included characteristic variables. We then turn to two methods of hedonic quality adjustment of non-comparable replacements items: *prediction* and *quality adjustment*. Both methods are based on using a base period hedonic regression to 'adjust' current period characteristics for quality changes. It is shown that the hedonic estimates used in each method are unbiased. As such their efficacy relies on the size of the *prediction intervals* attached to the predictions/adjustments and we compare these intervals under full and omitted variable specifications. In the empirical section we focus on the 'prediction method', and identify the extent of the prediction interval and how it changes over time and under parsimonious specifications.

The results in the following discussion hold more generally but for expositional purposes we consider only this simple example. Suppose that the correctly specified base period hedonic

regression is:

$$p_t(z_{1it}, z_{2it}) = \beta_{0t} + \beta_{1t} z_{1it} + \beta_{2t} z_{2it} + \varepsilon_{it} \qquad (1)$$

where p_t is the period t price of model i, z_{1it} and z_{2it} represent the value of its characteristics, and ε_{it} is an error term with the usual desirable properties. Estimating this equation using OLS yields $\hat{\beta}_{0t}$, $\hat{\beta}_{1t}$ and $\hat{\beta}_{2t}$ which are unbiased estimates of β_{0t}, β_{1t} and β_{2t}. Such hedonic regressions are regularly estimated by statistical agencies for "quality adjusting" the prices of missing goods.

Within this framework, we can also consider a hedonic equation where characteristic z_{2it} has been omitted,

$$p_t^*(z_{1it}, z_{2it}) = \beta_0^* + \beta_1^* z_{1it} + \varepsilon_{it}^* \tag{2}$$

If we estimate this regression, $\hat{\beta}_{0t}^*$ and $\hat{\beta}_{1t}^*$ are biased estimates of β_{0t} and β_{1t}. In particular,

$$\hat{\beta}_{0t}^* = \beta_{0t} + \beta_{2t}\hat{d}_{20t} + m_{0t}\varepsilon_t \tag{3a}$$

$$\hat{\beta}_{1t}^* = \beta_{1t} + \beta_{2t}\hat{d}_{21t} + m_{1t}\varepsilon_t \tag{3b}$$

where

$$m_{0t} = \frac{\sum (z_{1it} - \bar{z}_{1t})(\varepsilon_{it} - \bar{\varepsilon}_t)}{\sum (z_{1it} - \bar{z}_{1t})^2},$$

$$m_{1t} = \bar{\varepsilon}_t - \frac{\bar{z}_{1t} \sum (z_{1it} - \bar{z}_{1t})(\varepsilon_{it} - \bar{\varepsilon}_t)}{\sum (z_{1it} - \bar{z}_{1t})^2}$$

and

$$\hat{d}_{20t} = \bar{z}_{2t} - \hat{d}_{21t}\bar{z}_{1t}$$

$$\hat{d}_{21t} = \frac{\sum (z_{1it} - \bar{z}_{1t})(z_{2it} - \bar{z}_{2t})}{\sum (z_{1it} - \bar{z}_{1t})^2}$$

are the estimated coefficients from the auxiliary regression:

$$z_{2it} = d_{20t} + d_{21t} z_{1it} + \nu_{it}$$

where $E\nu_{it} = 0$, $E\nu_{it}^{\ 2} = \sigma_{2t}^2(1 - r_{12t}^2)$ and r_{12t} is the correlation coefficient between z_{1t} and z_{2t}. Taking expectations confirms that our coefficients are biased.

In order to make analytic comparisons of the predictive powers of (1) and (2) we need to know the distribution of ε_{it}^*. Since $p_{it}^* = p_{it}$,

$$\begin{aligned}
\varepsilon_{it}^* &= p_{it}^* - \beta_{0t}^* - \beta_{1t}^* x_{1it} \\
&= \beta_{0t} + \beta_{1t} x_{1it} + \beta_{2t} x_{2it} + \varepsilon_{it} - (\beta_{0t} + \beta_{2t} d_{20t}) - (\beta_{1t} + \beta_{2t} d_{21t}) x_{1it} \\
&= \varepsilon_{it} + \nu_{it}
\end{aligned}$$

Notice that when x_1 and x_2 are perfectly correlated, $\varepsilon_{it}^* = \varepsilon_{it}$.

A. Prediction

The UK Office of National Statistics (ONS) uses hedonic regressions to estimate missing back prices—Ball and Allen (2003) and Ball *et al*. (2004). To do so, they estimate the price of the non-comparable replacement item using fitted values from a hedonic regression. In doing so:

> "A wide range of attribute data is collected for both PCs and laptops. In the case of PCs, although much of the price is accounted for in "core" attributes such a processor speed

and memory size, changes in technology have led to attributes such as graphic and sound cards also having a significant influence. Much work is undertaken to ensure that the hedonic regressions do not suffer from missing-variable bias, so such things as on and off site warranty are also included." (Ball and Allen, 2003: 6).

It is well known that both our fully specified (1) and our omitted variables (2) hedonic regressions yield unbiased estimates of the price, p_t. For example, the predicted price of a good with characteristics $z_{1i\tau}$ and $z_{2i\tau}$ is:

$$\hat{p}_t(z_{1i\tau}, z_{2i\tau}) = \hat{\beta}_{0t} + \hat{\beta}_{1t} z_{1i\tau} + \hat{\beta}_{2t} z_{2i\tau}.$$

where $\tau > t$. Since $E(\hat{\beta}_{0t}) = \beta_{0t}$, $E(\hat{\beta}_{1t}) = \beta_{1t}$ and $E(\hat{\beta}_{2t}) = \beta_{2t}$ it is straightforward to conclude that $E(\hat{p}_t(z_{1i\tau}, z_{2i\tau})) = E(p_t(z_{1i\tau}, z_{2i\tau}))$. We can write the variance of the prediction error as:

$$\begin{aligned}
&\text{var}(\hat{p}_t(z_{1i\tau}, z_{2i\tau}) - p_t(z_{1i\tau}, z_{2i\tau})) = (z_{1i\tau} - \bar{z}_{1t})^2 \,\text{var}(\hat{\beta}_{1t}) + (z_{2i\tau} - \bar{z}_{2t})^2 \,\text{var}(\hat{\beta}_{2t}) \\
&+ 2(z_{1i\tau} - \bar{z}_{1t})(z_{2i\tau} - \bar{z}_{2t}) \,\text{cov}(\hat{\beta}_{1t}, \hat{\beta}_{2t}) + \sigma^2\left(1 + \frac{1}{n}\right) \qquad (4) \\
&= \frac{\sigma^2}{(1 - r_{12}^2)}\left(\frac{(z_{1i\tau} - \bar{z}_{1t})^2}{\sum (z_{1i\tau} - \bar{z}_{1t})^2} + \frac{(z_{2i\tau} - \bar{z}_{2t})^2}{\sum (z_{2i\tau} - \bar{z}_{2t})^2} - 2\frac{(z_{1i\tau} - \bar{z}_{1t})(z_{2i\tau} - \bar{z}_{2t}) r_{12}^2}{\sum (z_{1i\tau} - \bar{z}_{1t})(z_{2i\tau} - \bar{z}_{2t})} + \right) + \sigma^2\left(1 + \frac{1}{n}\right)
\end{aligned}$$

Now consider the predicted price using (2). It is also known that despite the fact that it is misspecified, using it for prediction still yields unbiased estimates of price. To illustrate, the predicted price in this case is

$$\begin{aligned}
&\hat{p}_t^*(z_{1i\tau}, z_{2i\tau}) = \hat{\beta}_{0t}^* + \hat{\beta}_{1t}^* z_{1i\tau} \\
&= \beta_{0t} + \beta_{2t}\hat{d}_{20t} + m_{0t}\varepsilon + (\beta_{1t} + \beta_{2t}\hat{d}_{21t} + m_{1t}\varepsilon) z_{1i\tau} \\
&= \beta_{0t} + \beta_{1t} z_{1i\tau} + \beta_{2t}(\hat{d}_{20t} + \hat{d}_{21t} z_{1i\tau}) + m_{0t}\varepsilon + m_{1t}\varepsilon z_{1i\tau} \\
&= \beta_{0t} + \beta_{1t} z_{1i\tau} + \beta_{2t}\hat{z}_{2i\tau} + m_{0t}\varepsilon + m_{1t}\varepsilon z_{1i\tau}
\end{aligned}$$

Taking expectations, we see that $E(\hat{p}_t^*(z_{1i\tau}, z_{2i\tau})) = E(p_t(z_{1i\tau}, z_{2i\tau}))$. We can similarly write the prediction error as:

$$\begin{aligned} &\text{var}(\hat{p}_t^*(z_{1i\tau}, z_{2i\tau}) - p_t(z_{1i\tau}, z_{2i\tau})) = (z_{1i\tau} - \bar{z}_{1t})^2 \,\text{var}(\hat{\beta}_{1t}^*) + \sigma^2\left(1 + \frac{1}{n}\right) \\ &= \sigma^{*2}\left(\frac{(z_{1i\tau} - \bar{z}_{1t})^2}{\sum (z_{1i\tau} - \bar{z}_{1t})^2}\right) + \sigma^2\left(1 + \frac{1}{n}\right) \end{aligned} \tag{5}$$

One might guess that the variance of the prediction error might be greater using the misspecified hedonic regression but it is not an obvious consequence and depends on the relative magnitudes of $\sigma_t^2/(1-r_{12t}^2)$ and $\sigma_t^2 + \sigma_{2t}^2(1-r_{12t}^2)$ and on the relative magnitudes of the first bracketed term in the final line of (4) and (5). It should be clear that if characteristics 1 and 2 are sufficiently correlated, the prediction error using the misspecified hedonic is less than the prediction error using the full hedonic. To evaluate the extent to which prediction error increases (if at all), one must turn to a practical application. We do so in Section III

B. Quality Adjustment

The US Bureau of Labor Statistics (BLS) uses the coefficients from hedonic regressions to estimate the value of the change in quality from one model to the next which is then used to "quality adjust" the back price. For example, using the fully specified hedonic (1), if the base period item has characteristics z_{1it} and z_{2it} and the current period item has characteristics $z_{1i\tau}$ and $z_{2i\tau}$, the BLS predicts the base period price of a good the characteristics of item i as:

$$\hat{p}_t^a(z_{1i\tau}, z_{2i\tau}) = p_t(z_{1it}, z_{2it}) + \hat{\beta}_{1t}(z_{1i\tau} - z_{1it}) + \hat{\beta}_{2t}(z_{2i\tau} - z_{2it})$$
$$= \beta_{0t} + \beta_{1t} z_{1it} + \beta_2 z_{2it} + \varepsilon_{it} + \hat{\beta}_{1t}(z_{1i\tau} - z_{1it}) + \hat{\beta}_{2t}(z_{2i\tau} - z_{2it})$$

Taking expectations, we see that this quality adjustment yields the base period expected price of a product with the characteristics of current period item i. Computing the variance of the prediction error, we get:

$$\begin{aligned}
&\text{var}(p_t^a(z_{1i\tau}, z_{2i\tau}) - p_t(z_{1i\tau}, z_{2i\tau})) = (z_{1i\tau} - z_{1it})^2 \,\text{var}(\hat{\beta}_{1t}) + (z_{2i\tau} - z_{2it})^2 \,\text{var}(\hat{\beta}_{2t}) \\
&+ 2(z_{1i\tau} - z_{1it})(z_{2i\tau} - z_{2it}) \,\text{cov}(\hat{\beta}_{1t}, \hat{\beta}_{2t}) + 2\sigma_t^2 \qquad (6) \\
&= \sigma_t^2 \left(\frac{(z_{1i\tau} - z_{1it})^2}{\sum (z_{1i\tau} - \bar{z}_{1t})^2 (1 - r_{12}^2)} + \frac{(z_{2i\tau} - z_{2it})^2}{\sum (z_{2i\tau} - \bar{z}_{2t})^2 (1 - r_{12}^2)} - 2\frac{(z_{1i\tau} - z_{1it})(z_{2i\tau} - z_{2it}) r_{12}^2}{\sum (z_{1i\tau} - \bar{z}_{1t})(z_{2i\tau} - \bar{z}_{2t})(1 - r_{12}^2)} \right) + 2\sigma_t^2
\end{aligned}$$

This is quite similar to, but differs from (4) in two ways. First, rather than deviations of the current period characteristic from the base period mean the quality adjustment method looks at deviations of the current period characteristic from the base period characteristic. Second, $2\sigma_t^2 > \sigma_t^2(1 + 1/n)$. If we expect $z_{1i\tau}$, z_{1it}, $z_{2i\tau}$, and z_{2it} to be randomly drawn from the same sample, the former will be twice that of the comparable terms from (4). Thus the variance of the error here will on average be strictly greater than variance from direct prediction and when the sample size is large, it will be nearly double that from direct prediction. However, there is some notion that substitute goods are chosen to minimize the change in observable characteristics. If this is the case then it is not clear which formula yields the lower variance.

As with traditional prediction, we can examine the properties of the "quality adjustment" predictor using the coefficients from our misspecified hedonic regression,

$$\hat{p}_t^{a*}(z_{1i\tau}, z_{2i\tau}) = p_t(z_{1it}, z_{2it}) + \hat{\beta}_{1t}^{*}(z_{1i\tau} - z_{1it})$$
$$= \beta_{0t} + \beta_{1t} z_{1it} + \beta_{2t} z_{2it} + \varepsilon_{it} + (\beta_{1t} + \beta_{2t}\hat{d}_{21t})(z_{1i\tau} - z_{1it})$$
$$= \beta_{0t} + \beta_{1t} z_{1i\tau} + \beta_{2t} z_{2it} + \beta_{2t}((\hat{z}_{2i\tau} - d_{20t}) - (\hat{z}_{2it} - d_{20t})) + \varepsilon_{it}$$
$$= \beta_{0t} + \beta_{1t} z_{1i\tau} + \beta_{2t} z_{2it} + \beta_{2t}(\hat{z}_{2i\tau} - \hat{z}_{2it}) + \varepsilon_{it}$$

Since $E(\hat{z}_{2it}) = z_{2it}$, $E(\hat{z}_{2i\tau}) = z_{2i\tau}$ and $E(\varepsilon_{it}) = 0$ this gives us unbiased estimates of the base period price of current period item i. The variance of this predictor is:

$$\text{var}(\hat{p}_t^{a*}(z_{1i\tau}, z_{2i\tau}) - p_t(z_{1i\tau}, z_{2i\tau})) = (z_{1i\tau} - z_{1it})^2 \,\text{var}(\hat{\beta}_{1t}^{*}) + 2\sigma_t^{\,2}$$
$$= \sigma_t^{*2}\left(\frac{(z_{1i\tau} - z_{1it})^2}{\sum (z_{1i\tau} - \bar{z}_{1t})^2}\right) + 2\sigma_t^{\,2} \qquad (7)$$

Similar to the prediction method, the variance of the prediction error may actually be smaller using the misspecified hedonic.

C. Effect on prediction out-of-sample

The standard error for predictions, which can be used to generate, say, 95% confidence intervals ($\pm 1.96 SE(\hat{p})$), is given by $SE(\hat{p})$:

$$SE(\hat{p}) = \sqrt{\sigma^2\left(1 + \frac{1}{n}\right) + \sum_{j=1}^{k}\sum_{l=1}^{k}\left(z_j^t - \bar{z}_j^t\right)\left(z_l^t - \bar{z}_l^t\right)\text{cov}\left(\hat{\beta}_j, \hat{\beta}_l\right)} \qquad (8)$$

where σ^2 is the standard error of the regression and x_{i0} is the distance out-of-sample observation 0 on variable *i* is from the mean of variable *i* and x_{j0} is the distance out-of-sample observation 0

on variable *j* is from the mean of variable *j*. The prediction interval 'bounds' on the prediction is determined by the fit of the regression model (better fit—smaller bounds), the sample size (larger sample size—smaller bounds), and the distance the out-of-sample observations on the explanatory variables are away from the respective means of the explanatory variables used in the regression (closer to the mean—smaller bounds). However, if quality improves over time the predictive interval will widen**.**

III. Data and Results

A. Data

The empirical work uses monthly 1998 'scanner' data for washing machines. In each month all the transactions from the bar-code readers (scanners) of retailers are aggregated for each model of the product in each of four outlet-types. This provides, for each model in each month sold in each outlet-type, unit values ('price'), sales volumes and sales values. The scanner data also include for each model a unique model identifier and a set of variables on quality characteristics. In 1998 the data set comprised 7,750 observations (models in an outlet-type in a month) made up from about 1.5 million transactions worth about £550 million. The quality characteristics for each product are given in Annex 1 and are listed in the first column of Table 1.

[Table 1 about here]

B. Results

Table 1 provides the results from an OLS hedonic regression estimated for January 1998 on the characteristics (including 'type' and 'build'), outlet-type and brands of washing machines. The first column includes results using all the variables. The estimates are reasonable with $\overline{R}^2 = 0.70$,[3] the model neither rejecting a null of homoskedasticity, nor Ramsey's RESET specification test and the signs generally according with *a priori* expectations, though the residuals are non-normal (Jarques-Bera).[4]

We then re-estimate the model omitting some sets of variables. We exclude in turn: (i) all brands; (ii) all minor brands[5]; (iii) outlet-types; (iv) dimensions (height, width, and depth); (v) variables with *t*-statistics on their coefficients in the full model of less than 1; (vi) variables with *t*-statistics of less than 2.

Bias from using coefficients

We start by considering the effects on included variables by omitting sets of variables. We established in (3) that bias might arise to the coefficients on included variables from omitting

[3] Previous work on washing machines (Silver and Heravi, 2003) restricted the OLS estimator to observations with sales of 30 or more so as to not unduly influence the estimates by observations with low sales and unusual pricing. In this work we do not wish such exploratory data analysis to influence the results and use all of the data. The $\overline{R}^2$ is lower than usual.

[4] This deviation from the normality assumption may not permit correct inferences to be drawn on the coefficients. However, a heteroskedasticity-consistent covariance matrix estimator (HCCME) was used following White (1980) to allow asymptotically correct tests to be undertaken. A wild bootstrap estimator is usually advised for estimating models with heteroskedastic and skewed residuals due to small-sample bias in the HCCME. Davidson and Flachaire (2001) show that the wild bootstrap is only necessary to alleviate *small-sample* bias; the HCCME estimator is appropriate for the large sample tests in this study.

[5] These are brands which in January 1998 accounted for less than 5% of sales value. The remaining six brands accounted for about two-thirds of sales value. Minor brands were grouped into a dummy variable 'other brands' with the coefficients for the major brands remaining as benchmarked on AEG for comparison with the full model.

variables. The extent of such bias to included coefficients would depend on the products of the coefficients of the omitted variables and the multicollinearity between omitted and included variables in auxiliary regressions. Given the number of coefficients in Table 1 we focus the discussion on only five variables. These were selected as the five variables with the highest *standardised* beta coefficients and conveniently include a type of washing machine (top loader), characteristic (spin-speed in 100 rpm), build (built-under and integrated), outlet-type (catalog), and brand (Hoover)—all in bold in Table 1. The coefficients for top loader (benchmarked on freestanding) are relatively similar across specifications except when variables on 'dimensions' are dropped, in which case the coefficient nearly halves. This implies that both the coefficients on the omitted dimension variables are relatively high as are the (conditional) coefficients on top-loader in auxiliary regressions of each dimension variable on top-loader, i.e. being a top-loader is related to the size of the machine, and its omission affects the estimates on top-loader, which makes sense.

In Table 1 the coefficient on 'spin-speed' hardly deviates across specifications of the regression implying that most brands, outlets–types, dimensions and types of machines have a variety of spin speeds. Only particular brands or outlet-types may make or sell built-under and integrated washing machines (BUI) and we note some variability in the BUI coefficients as variables on brand and outlet-type are omitted. The coefficients on sales of washing machines in 'catalog' outlets is robust to changes in the specification of the regression again reflecting no bias to brand or type of machine sold there. However, the coefficient of the brand 'Hoover' is very sensitive to the specification. Note the dramatic fall when outlet-types are excluded, with Hoover sales being (conditionally) biased to specific outlet-types. Thus individual coefficients

are sensitive to omitted variable bias in a way that can be determined in section II and, arising from this, can be predicted on *a priori* grounds.

Effect on predicted values in-sample

It is well known that a reduction in MSE could be obtained by omitting variables with (albeit estimated) coefficients less than unity. We undertook F-tests to compare the unrestricted sum of errors from the full specification with the restricted sum of squared errors from the parsimonious specifications and, Table 1, rejected the null of no difference for each parsimonious specification with the exception of dropping variable with t-statistics less than unity. We see from Table 1 that adopting a criterion of excluding variables with *t*-statistics less than unity decreases the sum of squared errors. Yet hedonic regressions estimated in January is considered here for predictions from a January regression in January. Our concern with quality adjusting the prices of non-comparable replacement models is with predictions from a January regression *in subsequent months*. We therefore examine out-of-sample predictions.

Effect on predictions out-of sample

Consider a hedonic function $\hat{p}_{it} = h_{it}(z_{it})$ where $\hat{p}_{it}$ is the predicted price of model *i* in period *t*, z_{it} is a vector of characteristics for model *i* in period *t* and $h_{it}(\cdot)$ a hedonic function estimated in period *t.* A price comparison requires an adjustment to one of the prices for the quality difference.

$$P(z_{it}) = \frac{h_{i\tau}(z_{i\tau})}{h_{it}(z_{i\tau})} = \frac{\hat{p}_{i\tau}}{h_{it}(z_{i\tau})} \approx \frac{p_{i\tau}}{h_{it}(z_{i\tau})} \quad (9)$$

$$P(z_{it}) = \frac{h_{i\tau}(z_{it})}{h_{it}(z_{it})} = \frac{h_{i\tau}(z_{it})}{\hat{p}_{it}} \approx \frac{h_{i\tau}(z_{it})}{p_{it}} \quad (10)$$

Both price comparisons measure the price change of a fixed set of characteristics, it is just that in (9) the characteristics are held constant in period τ while in (10) they are held constant in period *t*. Neither comparison can be held to be more appropriate, though (9) has a practical advantage. If the numerator in (9), $\hat{p}_{i\tau}$, is replaced by the actual price, $p_{i\tau}$, it can be seen that there is only a need to estimate a hedonic regression in period *t*, and not in real time.[6] Of course the predicted and actual prices may differ, but such price comparisons may be undertaken over many, or even all, observations and the expected mean predicted price may be quite close the mean actual price.[7]

We consider the omission of variables with regard to their predictive ability out-of-sample, in a future period as in the denominator in (9), for this is what they are to be used for. Given the importance of predictions holding out-of-sample, our concern is that any pruning of the regression equation will adversely effect the predictions in subsequent periods. We first

[6] Note that the coefficients will at some stage become out-of-date and the comparison of current period characteristics at base period characteristic prices less meaningful. Had comparisons of base period characteristics at current period prices also been evaluated this 'out-of-datedness' would be reflected in a Laspeyres-type to Paasche-type spread as explained in Silver and Heravi (2003a). The manner and pace with which coefficients can change is impressively recorded in van Mulligen (2003).

[7] They need not be equal. Suppose $ln\, p_j = a + bx_j + u_j$, then $ln\, \hat{p}_j = a + bx_j$

(in general, $p_j = exp(h(x_j) + u_j)$.

$$\left[\Pi_j^n p_j\right]^{1/n} = \left[\Pi_j^n \left(exp(h(x_j) + u_j)\right)\right]^{1/n} = \left[\Pi_j^n exp(h(x_j))\right]^{1/n} \left[\Pi_j^n exp(u_j)\right]^{1/n}$$

Since $\left[\Pi_j^n exp(u_j)\right]^{1/n} \neq 1$, then $\left[\Pi_j^n p_j\right]^{1/n} \neq \left[\Pi_j^n \hat{p}_j\right]^{1/n}$.

We calculated the mean of the residuals in each period and found minimal deviation from unity.

consider how well predictions from an all variables specification in a base period hedonic regression perform when they predict out-of-sample current period sample data. We do this to identify any decline in the predictive ability as the current period is further removed from the reference hedonic period. We then compare the result with omitted variable specifications.

The 'all variables' specification

[Table 2 about here]

Table 2 provides the results for the plus/minus error margin of the prediction interval for, in turn, February, March… December using, in each case, a hedonic regression estimated in January. The first set of results is based on a hedonic regression estimated for January that includes *all* the variables. The prediction interval varies across observations so the entries are for the means and standard deviations of the individual error margins; the $\pm 1.96 \times SE(\hat{p})$ given in equation (8). For February 98, based on a January 1998 hedonic regression, we would expect 95% of predictions to fall within $\pm 44\%$ of the mean price.[8] The mean prediction interval is relatively large which may be expected from the relatively poor fit reflected in an $\overline{R}^2 = 0.704$ in Table 1.[9] In fact such out-of-sample prediction intervals provide a more salient means by which we can judge what we mean by a 'satisfactory' fit.

Of course there was some variability in the prediction intervals underlying this mean of the 425 bounds. However, Table 2 finds the dispersion to be relatively small with a standard deviation

[8] Bear in mind the use of percentage intervals from a semi-logarithmic formulation.
[9] We note again previous studies have higher values our work here following fairly basic practice of an OLS regression using all of the data, rather than volume cut-offs to exclude unusual sales (Silver and Heravi, 2003).

of about 2%. The driving force behind equation (8) is the standard error of the regression, σ, which is a constant for each interval estimate, thus explaining the relatively low dispersion.

We would expect the prediction interval to 'fan out', as the differences between the *z*'s in subsequent months diverge from the January 1998 means. But, unexpectedly, they do not. The width of the interval is also determined (equation (8) by the variation the explanatory variable *z*'s away from the *z*'s in January. As each quality variable in a current month departs from its mean in January, the prediction bounds should 'fan' out. While there is an increasing trend to the error margin in Table 2, it is relatively small. In this multivariate context the changes in some variables must be given more weight than others. The prediction interval is determined by the product of the variance of the estimated coefficients and the variance of the observations from their means in January. Thus a given variable's impact on the interval will in part be due to the extent to which it changes over time and also the variance of its coefficient which will in turn be in part determined by any multicollinearity between it and other explanatory variables. It thus becomes difficult *a priori* to determine which variables to include to minimise the interval and the extent to which the errors will change over time.

The omitted variables specifications

Next Table 2 shows how the predictive interval changes as we omit variables. Dropping brands leads to a substantial increase in the 95% error margins, from 44% to 53% in December. There is a naturally smaller loss if only the major 5 brands are included, though the 95% error margins still have a notable increase to 47.5% in December. Dropping outlet-types also matters though, dropping dimensions less so with a December 95% error margin of 46.8%. The mean 95% error

margin is, for practical purposes, unaffected when variables with *t*-statistics less than unity are dropped. Of further note is that the standard deviation also decreases, as expected from the discussion in section 2. The dropping of variables with *t*-statistics less than 2 involved a fall in the number of explanatory variables from 34 to 25. The increase in the width of the 95% predictive interval was minimal, from plus/minus 44.2% to 45.5% in December, and the standard deviation of the intervals fell by half. This empirical exercise shows that such deletion is well worth while when there is an administrative burden to variable collection and use.

IV. Conclusions

Statistical agencies endeavour to estimate the best hedonic regression for use in quality adjusting their price indices. This effort may not be entirely necessary as quality adjustment using hedonics with omitted variables may yield comparable results. In particular, the dropping of regressors yields biased coefficient estimates, however, price predictions and quality adjustments still yield unbiased price estimates. Moreover, the variance of prediction errors may, depending on the application, actually be lower with some variables omitted. Thus there are efficiency gains to be had by estimating more parsimonious hedonic regressions.

We then consider this in an actual application using British scanner data on washing machines. We estimate hedonic regressions using specifications varying in their parsimony. As expected, under some specifications result in significant changes in estimated coefficients while under specifications estimated coefficients are relatively stable. Computing prediction intervals shows wide dispersion (±44% for the full specification), due to our relatively low initial $\overline{R}^2$ of about

0.70. However, these intervals do not appear to fan out for prediction characteristics further in the future. Even at its worst when all brand names have been dropped, the prediction interval increases to only to ±53%.

These results are still preliminary. At this point, our empirical results regarding prediction intervals is focused solely on prediction rather than "quality adjustment" as is done at the BLS. To be able to make some comparisons between the two methods, we will calculate prediction intervals for quality adjustment as well. Related to this point, the prediction intervals we calculate are using all observations rather than just the unmatched ones. Realistically, our prediction intervals should be computed using only unmatched observations.

In addition, rather than arbitrarily dropping particular groups of variables, we will drop/group variables in a stepwise manner using a criterion such as R^2 or MSE. For example, in the first step, it may be that when the "Neff" and "Tbend" brands are grouped together, R^2 falls the least in comparison to other possible droppings/groupings. In the second stage, it may be that the "load" characteristic is dropped, etc.

Annex 1 – Variable set on characteristics of washing machines: (i) Manufacturer (make) – dummy variables for about 20 makes; (ii) types of machine: 5 types – top-loader; twin tub; washing machine (WM); washer dryer (WD) with and without computer; WD with/without condensers; (iii) drying capacity of WD; (iv) height of machine in cms.; (v) width; (vi) spin speeds (100 rpm); 5 main ones are 800 rpm, 1000 rpm, 1100 rpm, 1200 rpm and 1400 rpm; (vii) water consumption; (viii) load capacity; (x) free standing, built-under and integrated; built-

under not integrated; built-in and integrated; (xi) vintage; (xii) outlet-type: chain multiple stores, mass merchandisers (department stores), independents, catalogues.

References

Adams, W.J. and Yellen, J.L. (1976). Commodity Bundling and the Burden of Monopoly *Quarterly Journal of Economics*, Vol. 90 (1976), pp. 475–498.

Armstrong, M. (1996). Multiproduct Nonlinear Pricing. *Econometrica*. Vol. 64, pp. 51-76.

Armstrong, M. (1999). Price Discrimination by a Many-Product Firm. *Review of Economic Studies*. Vol. 66, pp. 151-168.

Ball, A. and Allen, A. (2003). The Introduction of Hedonic Regression Techniques for the quality adjustment of computing equipment in the Producer Prices Index (PPI) and Harmonised Index of Consumer Prices (HICP), *Economic Trends*, 592, February, London: Office for National Statistics.

Ball, A., Waldron, K., Smith, K. and Hughes, J. (2004). Changes to Methodology Employed in the CPI and RPI from February 2004, *Economic Trends*, 604, March, London: Office for National Statistics.

Davidson, R. and Flachaire, E. (2001), "The Wild Bootstrap, Tamed at Last", *Institute of Economic Research Working Paper Series* no. 1000, Department of Economics, Queen's University, Ontario.

McAfee, R.P., McMillan, J. and Whinston, M. (1989). Commodity Bundling by a Monopolist, *Quarterly Journal of Economics*, May, 371-83.

Mirman, L.J. and Sibley, D. (1980). Optimal Nonlinear Prices for Multiproduct Monopolies, *The Bell Journal of Economics*, Fall, 1980 (with David Sibley).

Silver, M. and Heravi, S. (2003), The Measurement of Quality-Adjusted Price Changes. In Mathew Shapiro and Rob Feenstra (eds.), *Scanner Data and Price Indexes*, National Bureau of Economic Research, Studies in Income and Wealth, vol. 61, Chicago: University of Chicago Press, 277-317.

Silver, M. and Heravi, S. (2003a), On the Stability of Hedonic Coefficients and their Implications for Quality-Adjusted Price Change Measurement. Paper presented at the NBER-CRIW conference on Measurement Issues in Economics - The Paths Ahead, Essays in Honour of Zvi Griliches, Bethesda, MD, September 2003. http://www.nber.org/~confer/2003/CRIWf03/silver.pdf

van Mulligen, P.H. (2003). Quality Aspects in Price Indices and International Comparisons: Application of the Hedonic Method. Voorburg, Netherlands: Statistics Netherlands .

Table 1, Regression results for different specification

	All variables		Drop brands		Drop minor brands		Drop outlet-types		Drop dimensions		Drop t<1		Drop t<2	
	Coefs	**t-stats**	**Coefs**	**t-stats**	**Coefs**	**t-stats**	**Coefs**	**t-stats**	**Coefs**	**t-stats**	**Coefs**	**t-stats**	**Coefs**	**t-stats**
constant	3.068	2.33	4.491	3.48	1.864	1.08	3.126	2.45	6.605	7.84	3.138	2.49	5.687	14.58
Type (benchmarked 'washing machine (WM)')														
Top loader	**0.988**	8.13	**0.935**	8.15	**1.039**	7.76	**1.006**	8.08	**0.583**	4.11	**0.990**	9.19	**0.869**	7.14
Twin tub	-1.303	-7.61	-1.386	-8.08	-1.147	-6.67	-1.245	-7.37	-1.328	-6.63	-1.335	-7.60	-1.511	-9.41
Washer-Dryer (WD)	0.230	6.01	0.264	4.93	0.221	4.67	0.221	5.29	0.212	5.95	0.228	6.32	0.265	10.36
WM with chip	0.079	1.48	0.425	6.72	0.170	2.59	0.050	0.68	0.079	1.60	0.097	1.76		
WD with chip	0.396	5.68	0.487	5.13	0.429	5.34	0.371	4.87	0.389	5.00	0.394	5.87	0.467	8.07
Characteristics														
Height	-0.007	-2.53	-0.008	-2.55	-0.008	-2.12	-0.006	-2.50			-0.006	-2.71	-0.008	-3.73
Depth	0.014	1.63	0.021	2.82	0.020	2.10	0.012	1.44			0.014	1.71		
Width	0.030	5.73	0.031	5.59	0.034	5.60	0.032	5.86			0.031	6.19	0.017	2.96
Load	0.002	0.47	-0.005	-1.24	0.006	1.20	0.003	0.72	-0.001	-0.22				
Spin speed (100)	**0.054**	7.10	**0.053**	7.16	**0.053**	7.10	**0.050**	6.79	**0.058**	7.36	**0.053**	7.39	**0.055**	7.88
Water	-0.006	-6.76	-0.007	-7.63	-0.005	-7.09	-0.006	-6.38	-0.006	-6.64	-0.006	-8.87	-0.007	-9.44
Conditioner	0.078	1.49	0.037	0.20	0.101	1.74	0.079	1.40	0.075	1.43	0.060	1.57		
Vintage	0.011	1.44	0.001	-0.36	0.017	2.01	0.010	1.40	-0.005	-0.70	0.014	1.44		
Build (benchmarked 'freestanding)														
Built-under & integ.	**0.595**	14.76	**0.555**	8.64	**0.604**	12.90	**0.567**	14.15	**0.586**	14.54	**0.568**	14.12	**0.545**	12.93
Built-under not integ.	0.104	0.77	-0.160	-2.59	-0.012	-0.62	0.137	0.83	0.085	0.69				
Built-in & integrated	0.945	16.87	0.873	5.29	0.888	11.55	0.948	17.22	0.962	16.81	0.945	19.96	0.998	27.69
Outlet type (benchmarked on 'chain multiple)														
Mass merchandiser	0.081	3.03	0.088	2.95	0.082	2.94			0.080	2.94	0.082	3.08	0.068	2.94
Independent	0.071	2.28	0.104	2.66	0.085	2.54			0.079	2.20	0.071	2.28	0.066	1.97
Catalogue	**0.232**	8.76	**0.223**	7.59	**0.225**	8.11			**0.228**	8.30	**0.233**	8.76	**0.214**	8.67
Brand (benchmarked on 'AEG')														
Siemens	0.225	4.10					0.191	3.22	0.224	3.99	0.225	4.34	0.224	4.62
Bauke	0.054	0.75					-0.001	-0.52	0.090	1.91				
Hoover	**-0.127**	-3.75			**-0.150**	-3.86	**-0.091**	-2.68	**-0.118**	-3.74	**-0.127**	-4.22	**-0.085**	-3.74
Bosch	0.167	4.15					0.184	4.46	0.182	4.69	0.161	4.45	0.207	7.13
Miele	0.552	8.37			0.448	5.65	0.569	9.00	0.583	10.75	0.539	8.20	0.653	19.55

Table 1 continued

Candy	-0.106	-2.59					-0.130	-3.00	-0.175	-4.56	-0.113	-2.74	-0.119	-3.55
Ariston	-0.053	-1.66					-0.013	-0.72	-0.089	-2.73	-0.050	-1.82	-0.071	-2.85
Zannusi	0.165	3.71		0.169	3.68		0.207	4.43	0.143	2.81	0.156	4.60	0.171	6.85
Electric	0.068	0.47					0.104	0.89	0.031	-0.06				
Indesit	-0.137	-2.56		-0.144	-2.73		-0.119	-2.21	-0.194	-3.43	-0.133	-2.66	-0.151	-3.07
Neff	-0.102	-2.64					-0.057	-2.04	-0.106	-3.11	-0.109	-3.21	-0.061	-2.05
Credo	-0.011	-0.72					0.006	-0.38	-0.029	-1.12				
Tbend	-0.118	-3.12		-0.128	-3.33		-0.119	-3.06	-0.122	-3.15	-0.120	-3.32	-0.098	-2.68
Hotpoint	-0.055	-1.76		-0.093	-2.61		-0.025	-1.04	-0.011	-0.76	-0.048	-1.82		
Servis	-0.204	-4.00					-0.170	-3.25	-0.269	-5.19	-0.204	-4.18	-0.255	-6.00
Other brands				-0.014	-0.90									
Diagnostic statistics														
Rbar-squared	0.704		0.604	0.68		0.651		0.604		0.707		0.694		
n - observations	420		420	420		420		431		420		431		
k - variables	34		19	25		31		31		30		25		
Sum sqrd errors	12.30		17.54	14.08		16.62		14.36		12.34		13.85		
Std error of regress	0.181		0.210	0.189		0.197		0.192		0.180		0.187		
Ramsey's RESET	0.155		4.256*	5.64*		0.001		0.317		0.288		0.678		
L.M. het test	2.435		17.13	0.68*		1.790		2.926		2.335		3.027		
Jarque-Bera	6216***		1431***	3568***		4602		4691***		6171***		5224***		
F=test; null all coefs=	23.68***		32.95***	33.44***		20.10***		22.00***		23.99***		29.68***		
F-test: null omitted coefs=0.			10.92	6.16***		45.04***		21.39***		0.29		5.38***		

Table 2, Summary measures of prediction intervals

		All variables			Drop brands			Drop minor brands		
	Number of observations	Means of predicted prices, £	Prediction interval 95% standard error Mean	Std. dev.	Means of predicted prices, £	Prediction interval 95% standard error Mean	Std. dev.	Means of predicted prices, £	Prediction interval 95% standard error Mean	Std. dev.
February	425	397.299	1.438	1.026	395.176	1.525	1.027	397.943	1.470	1.036
March	531	403.877	1.439	1.026	401.754	1.527	1.027	404.411	1.471	1.036
April	535	408.581	1.440	1.032	405.105	1.528	1.034	408.192	1.474	1.044
May	565	414.665	1.439	1.017	408.886	1.525	1.016	414.610	1.471	1.021
June	620	419.578	1.440	1.026	411.135	1.527	1.026	419.848	1.473	1.035
July	652	421.403	1.441	1.028	410.459	1.527	1.030	421.279	1.474	1.039
August	628	425.434	1.440	1.024	414.990	1.527	1.025	426.089	1.473	1.033
September	631	429.620	1.442	1.030	416.301	1.529	1.032	430.193	1.475	1.041
October	618	428.444	1.443	1.032	417.026	1.530	1.033	430.057	1.476	1.043
November	574	426.614	1.441	1.025	415.843	1.528	1.026	429.158	1.474	1.034
December	591	427.028	1.442	1.027	415.823	1.528	1.027	429.707	1.475	1.036

		Drop dimensions			Drop t<1			Drop t<2		
		Means of predicted prices, £	Prediction interval 95% standard error Mean	Std. dev.	Means of predicted prices, £	Prediction interval 95% standard error Mean	Std. dev.	Means of predicted prices, £	Prediction interval 95% standard error Mean	Std. dev.
February		397.588	1.466	1.012	397.394	1.440	1.022	398.102	1.455	1.016
March		404.733	1.467	1.012	403.840	1.440	1.023	404.685	1.455	1.015
April		408.135	1.467	1.013	408.489	1.442	1.027	408.445	1.455	1.019
May		414.630	1.466	1.011	414.583	1.440	1.015	416.059	1.454	1.007
June		418.162	1.467	1.012	419.636	1.442	1.023	420.048	1.454	1.013
July		419.129	1.467	1.013	421.615	1.442	1.024	421.612	1.455	1.018
August		422.589	1.467	1.013	425.625	1.442	1.021	425.807	1.454	1.013
September		425.820	1.468	1.014	429.895	1.443	1.025	429.393	1.456	1.018
October		423.959	1.468	1.014	428.898	1.444	1.027	427.263	1.456	1.018
November		421.512	1.467	1.013	426.977	1.443	1.022	425.602	1.455	1.014
December		421.751	1.468	1.013	427.447	1.443	1.023	425.912	1.455	1.014