1352.0.55.099

Australian Bureau of Statistics

Research Paper

Reviewing the ABS' Hedonic Regression Model for Desktop Computers

www.abs.gov.au

Research Paper

Reviewing the ABS' Hedonic Regression Model for Desktop Computers

Charity Liaw and Steve Lane

Analytical Services Branch

Methodology Advisory Committee

7 November 2008, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 25 MAR 2009

ABS Catalogue no. 1352.0.55.099

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Charity Liaw, Analytical Services Branch on Canberra (02) 6252 5578 or email <analytical.services@abs.gov.au>.

CONTENTS

APPENDIXES

ABBREVIATIONS

ABS	Australian Bureau of Statistics
APMI	Price Index of Articles Produced by Manufacturing Industries
ANZSIC	Australian and New Zealand Standard Industrial Classification
ASB	Analytical Services Branch
ILO	International Labour Organization
IMF	International Monetary Fund
OECD	Organisation for Economic Co-operation and Development
PPI	Producer Price Indexes
MAC	Methodology Advisory Committee
US	United States of America

REVIEWING THE ABS' HEDONIC REGRESSION MODEL FOR DESKTOP COMPUTERS

Charity Liaw and Steve Lane
Analytical Services Branch

ABSTRACT

Quality change has long been recognised as perhaps the most serious measurement problem in estimating price indexes. When faced with the task of measuring prices for products that undergo rapid quality change (for example, consumer durables such as computers, whitegoods and cars), international best practice is to develop hedonic price indexes, provided suitable source data are available (Australian Bureau of Statistics, 2005).

In 2003, the Australian Bureau of Statistics (ABS) introduced a hedonic price index for desktop computers into the Producer Price Indexes (PPI). In 2008, a review of a part of method used to construct the index was undertaken, to ensure that the index remains relevant, given the fast evolving nature of computer technology.

This paper first details the review of the desktop computer price index, and then discusses how the recommendations arising from the review could be extended to provide a framework for the construction of price indexes for other consumer durables.

1. INTRODUCTION

Price indexes must measure pure price change and therefore must account for quality change. When pricing a good which undergoes rapid quality change, international agencies such as the Organisation for Economic Cooperation and Development (OECD), the International Labour Organization (ILO) and the International Monetary Fund (IMF) advocate the use of hedonic methods.

A hedonic price index is any price index that utilises, in some manner, a hedonic function – a function (often a regression model) which relates the price of a good to its attributes. The case for using hedonic price indexes is particularly strong for goods which are sold in the form of models (e.g. Manufacturer number 123456) characterised by attributes (e.g. *Vendor* = "IBM", *CPU Type* = "Pentium", *CPU Speed* = "3000", …).

The ABS began investigating the feasibility of developing a hedonic price index of desktop computers in 2001. The ABS required such an index to have the following properties (ABS, 2005):

1. The price index must measure pure price change; in the context of personal computers, this means that the price index must correctly account for rapid changes in the capability of personal computers.
2. The price index must be representative; that is, the index should be representative of transactions occurring within the Australian computer market, and must account for both the price changes of continuing models of personal computers, and those price changes associated with new models with new and/or improved characteristics entering the marketplace.
3. Movements in the price index should be easily explainable to users. Specifically, the period-to-period movements in the price index must be able to be decomposed to show the impact of the price changes for continuing models of personal computers, and the impact of the introduction of new models (with new and/or improved characteristics) to the marketplace.
4. The price index must be an improvement over existing methods that utilise data from United States of America (US) statistical agencies. In particular, the price index must avoid the counterintuitive movements that can be observed at times when using US data adjusted by $US/$A exchange rates series.

In May 2003, Jan de Haan of Statistics Netherlands presented a new hedonic technique to the Inter-Secretariat Working Group on Price Statistics (de Haan, 2003). This method, henceforth referred to as the 'de Haan double imputation method' satisfied all the requirements outlined above and was hence used by the ABS to create a price index for desktop computers. This price index was introduced into the Producer Price Indexes in 2003 (ABS, 2005) and into the Consumer Price Index and the National Accounts in 2005 (Purcell, Esguerra and Branson, 2007).

As Appendix A explains, the de Haan double imputation method calculates price changes for matched models from observed price changes and imputes price changes for unmatched models (i.e. superseded and new models) from a hedonic function (i.e. from a regression model).

In 2008, a review of the regression model used in the construction of the ABS' de Haan double imputation hedonic price index for desktop computers was undertaken by the Analytical Services Branch (ASB) in the ABS. A paper detailing the review and its results was presented to the Methodology Advisory Committee (MAC) in November 2008. This paper is an updated version of that original paper, which incorporates the many valuable suggestions made by the committee.

Sections 2 to 4 of this paper discuss the review, whilst Section 5 discusses how its recommendations could be extended to provide a framework for the construction of price indexes for other consumer durables.

2. CURRENT METHODOLOGY

This section describes the price imputation process currently used by the Producer Price Indexes Section in the ABS, the rationale behind it, and its identified shortcomings.

2.1 The process

Under the de Haan double imputation method, price changes for unmatched models are imputed using a regression model. The ABS imputes price changes for unmatched desktop computer models as follows.

Data on the prices and attributes of desktop computers are collected manually from some retailer websites. Data from two consecutive periods (May 2008 and June 2008, for example) are pooled together.

Using weighted least squares regression, a double-log regression model is fitted to the data – i.e. an equation is estimated which expresses the natural logarithm of the price of a desktop computer as a linear function of:

- dummy variables which describe the categorical attributes of the computer;
- natural logarithms (or, in some cases, base two logarithms) of the continuous attributes of the computer;
- two-way interactions between selected attributes of the computer;
- three-way interactions between selected attributes of the computer;
- a time dummy, which is '1' if the observation is from the second pooled month, and '0' otherwise (e.g. if the data are from May 2008 and June 2008, the time dummy is '1' for all observations taken in June 2008, and '0' for all observations taken in May 2008); and
- an error term.

Many attributes and interactions are initially included in the regression model and those that do not have a significant relationship with price are removed using the backward selection method. The first five explanatory variables entered by the compiler, however, are forced into the regression model.

The coefficient of the time dummy is used to impute an index number for the change in the price of unmatched models.

2.2 The history of the regression model

In 2001, ASB investigated the application of hedonic regression to desktop computer data. Given data (sourced from the International Data Corporation) for April, June, July, August, September, November and December 2000, the study examined various options for the regression model, taking into account the following criteria:

- prior industry knowledge (i.e. an understanding of the importance of each attribute of a computer, and of how attributes are related to one another);
- results from the application of the Box–Cox methodology – this involves using the Box–Cox transformation to let the data determine what functional form is most appropriate;
- ease of interpretation of results; and
- simplicity of estimation.

The study concluded that, given the data used, "the double-log functional form, where only RAM and Cache appear in logarithmic form with a base of two, was the most optimal [regression] model" (Lim and McKenzie, 2001).

In 2003, the ABS introduced a de Haan double imputation hedonic price index for desktop computers into the PPI Articles Produced by Manufacturing Index (APMI). The regression model used in the construction of that index, informed by the 2001 investigation, changed over the years – for example, in 2006, changes were made because, from October 2005, R^2 values fell below the values expected (Turnbull, 2006).

2.3 Issues with the current methodology

As mentioned above, a double-log functional form was chosen as a result of an investigation conducted in 2001. Since 2001, the market for desktop computers has changed substantially – for example, as multi-core processors have become available, many consumers are choosing to buy multi-core processors rather than single-core processors, perhaps because multi-core processors generally consume less power (as each processor runs at a lower speed compared with a single-core processor) and output less heat.

A double-log functional form may be preferable if continuous explanatory variables have a large range but are 'bunched' within months. For example, a double-log form may have been preferable when processor speeds were increasing exponentially. Today, we are seeing a trend towards increasing numbers of processor cores rather than increasing processor speeds, thus a semi-log functional form may be preferable.

Briefly, using *RAM* as the characteristic in question, the semi-log functional form translates to a constant price (increase) per unit of *RAM* regardless of the overall size of *RAM*, whereas the double-log functional form exhibits a higher price increase when increasing *RAM* from lower levels than increasing *RAM* from higher levels. Thus, the double-log form provides 'insurance' should large changes in computers' characteristics occur in the future.

It is not clear whether the long-term view (which would support the double-log functional form) or the short-term view (which supports the semi-log functional form) is more appropriate, given that the aim is to measure price changes between successive months.

Recent changes in technology also lead one to believe that the attributes driving the prices of desktop computers today are different from those driving prices in 2001. Data on 'new' attributes, such as video cards, are not available, thus the current model may be incorrectly specified, leading to outliers (which are currently dropped) and lowered predictive power. Data on now irrelevant attributes are still being collected.

The data are collected manually, from retailer websites. As well as being extremely labour intensive, the manual process is prone to errors.

Questions for MAC

- Technological change causes drastic and sudden changes in the markets for consumer durables (most markedly, in the markets for electronics such as computers, televisions and DVD players). New categories of attributes appear (e.g. new *CPU Types*), and entirely new attributes appear (e.g. video cards). Furthermore, new attributes can change the relationships between price and existing attributes (e.g. the introduction of multi-core processors changed the relationship between *Price* and *CPU Speed* – it is now not valid to include *CPU Speed* without including an interaction between *CPU Speed* and the number of processor cores). Given cost barriers (it is not feasible to collect all attributes which may, perhaps, influence price; and also not feasible to collect data retrospectively), what is the best way to deal with this problem?

Response from MAC

- The difficulty of this problem was acknowledged. However, no strategies for dealing with this problem were identified.

The inclusion of many of the interactions in the current model seems to be unnecessary. The inclusion of an interaction between two explanatory variables is defensible only when we have reason to believe that one of those variables changes the effect, on the response variable, of the other explanatory variable. In addition, higher order interactions implicitly introduce multicollinearity into a regression model, which can inflate estimates of the variation of coefficients quite severely. In the current model, half of the included explanatory variables have a Variance Inflation Factor (a measure of the severity of multicollinearity) many times larger than the rule-of-thumb cut-off of 10, indicating substantial redundancy.

Conceptually, squared terms can be viewed in the same way as interactions. If there is no theoretical reason to assume a quadratic structure, nor evidence of a quadratic structure in the data (e.g. a quadratic pattern in a plot of explanatory variable vs response variable, or explanatory variable vs residuals), the inclusion of squared terms is not defensible.

Regression modelling theory suggests that weights should not be used if the variable by which the data are grouped is an included explanatory variable. Weights are used in the current model, even though the data are grouped by *Vendor*, which is an included explanatory variable.

Terms should not be forced into a regression model without strong justification. Forcing the time dummy into the regression model may be justifiable, as its coefficient is a required output. Forcing the variable *Vendor* into the regression model may also be justifiable, as it is akin to a stratification variable, and thus its inclusion removes the need to complicate the model by adding weights. There seems to be weaker justification for forcing in other explanatory variables.

Questions for MAC

- Is it valid to force the time dummy into the regression model?
- Is it valid to force the stratification variable into the regression model, given that it is significant in most months?
- If the stratification variable is removed in the backward selection process, is it fair to assume that the design is non-informative, and thus we do not need to use weights?

Response from MAC

- Because the time dummy is the variable of interest, it should be forced into the model.
- Statistical significance should not entirely determine the included explanatory variables (because, for example, the functional form could be wrong). Variables which are believed to strongly influence price should be forced into the model. *Vendor* is one such variable. *Website* (the website from which the observation is collected) and *Brand–Model* (the 'name' of the model, e.g. "HP Presario") may be other such variables.
- Weights should be used – expenditure share weights are preferable. (An alternative is to sample proportional to expenditure share.) It was suggested that the weights used in this study, which were calculated in 2000, be updated.

As Appendix A explains, the data collected in two consecutive months are pooled, a regression is then run on that pooled data. Thus, models which exist in both months (i.e. matched models) appear in the regression dataset twice, making it likely that the regression assumption of independently and identically distributed errors will be violated.

Questions for MAC

- Is it valid to allow matched models to appear in the regression dataset twice, with no allowance for dependence?
- If not, how should matched models be dealt with, given that each pooled dataset contains approximately 800 observations, approximately 70% of which are matched models?

Response from MAC

- It is valid to allow matched models to appear in the regression dataset twice, with no allowance for dependence, as the dependence that exists is unlikely to be strong enough to be problematic.

At present, the following regression diagnostics are used:

- R^2 (the proportion of the variation in the dependent variable which is explained by the explanatory variables);
- plots of standardised residuals vs continuous variables; and
- plots of standardised residuals and various influence measures vs observation number.

Theory, however, suggests that the most important regression diagnostic is a plot of standardised residuals vs predicted values. Theory also suggests the use of Adjusted-R^2 to indicate predictive power, and the use of Variance Inflation Factors to indicate if multicollinearity may be a problem.

Checking that the signs of regression coefficients are consistent with the expected theoretical effects of the explanatory variables was used as a goodness-of-fit criterion by Lim and McKenzie (2001), and indeed by a number of other researchers in hedonic theory. However, Pakes (2003) suggests that after allowing for markups (which depend on endogenous and exogenous variables), "the hedonic regression is in a 'reduced form', i.e. its coefficients have no obvious interpretation in terms of economic primitives".

Questions for MAC

- Assuming competitive markets, Pakes' suggestion appears inconsistent with economic theory. To what extent is it important for coefficients to be consistent with theoretical effects?
    - In general?
    - In hedonic regression models for consumer durables?
    - In hedonic regression models for desktop computers?

Response from MAC

- The degree of markups is unclear, thus one cannot become overly concerned with this issue.

3. A PROPOSED CHANGE

As discussed in Section 2.3, computer technology is constantly, and rapidly, changing. Thus, we expect that the 'optimal' hedonic regression model for desktop computers will change over time. We propose a yearly review of the regression model, involving the following steps.

1. Consult IT professionals. "Knowing your product" (Triplett, 2000, quoted by Lim and McKenzie, 2001) is an essential part of any hedonic regression fitting process. As computing is a highly specialised field, price index practitioners must liaise with IT professionals, in order to garner ideas regarding which explanatory variables should (and should not) be included in the regression model, and what functional form may be appropriate. Hypotheses formed from those ideas must, however, be backed by exploratory data analysis.

2. Fit some theoretically appropriate models and produce, for each model considered, the following diagnostics:

    - R^2 and Adjusted-R^2;
    - a plot of standardised residuals vs predicted values;
    - plots of standardised residuals v.s leverage; and
    - Variance Inflation Factors.

3. Choose the 'optimal' model, taking the following into consideration:

    - the 'reasonableness' of the regression model, given the industry knowledge gained in (1.);
    - the diagnostics produced in (2.);
    - ease of interpretation and simplicity; and
    - the 'reasonableness' of the index produced using the regression model.

Questions for MAC

- What other factors should be taken into consideration when choosing an 'optimal' model?

Response from MAC

- Economic theory, which supports double-log and semi-log functional forms, should also be considered when evaluating the 'reasonableness' of the regression model (NB 'flexible' functional forms could also used).
- When a double-log or semi-log functional form is used, a bias correction should be added to the time dummy coefficient (see Appendix A for details).

4. THE 2008 REVIEW

The model currently used (described in Section 2.1) was reviewed by ASB, using the process described in Section 3.1. The results of that review are presented in this section.

4.1 Acquiring advice from IT experts

Consultations with the IT Client Service Section in the ABS, regarding the attributes that currently drive desktop computer prices, point to the following as the four most important attributes, in order of priority:

- *CPU Type*;
- *RAM Size*;
- *Video Card*; and
- *Monitor*.

In addition, the following suggestions were put forward:

- *Hard Drive Size* may be quite important from a consumer's perspective, as consumers can easily compare the hard drive size of one computer to the hard drive size of another.
- The interaction between *CPU Type* and *RAM* should be included because, together, these are a proxy for *CPU Model*.
- If *CPU Speed* is included, the interaction with *CPU Type* must be included because the number of processor cores affects the effect on price of a change in *CPU Speed*.
- Components tend to be packaged such that high-end processors are sold with high-end other components. This may look like an interaction, but, from an 'effect on price' point of view, it is not.
- The relationship between *Price* and *CPU Speed* is nonlinear (especially since *CPU Speed* can be a proxy for *CPU Model*).
- The relationship between *Price* and *Operating System* is highly linear.
- The relationship between *Price* and *Monitor Size* may be nonlinear.

- The following collected variables are unlikely to be important:
  - *Ethernet*;
  - *USB Ports* (although USB1.0 vs USB2.0 may be important during change-over to USB2.0); and
  - *Hard Drive Type* (although SATA1 vs SATA2 may be important during change-over to SATA2).
- We should consider collecting the following variables:
  - *CPU Model* – for Intel processors, this describes *CPU Type*, *CPU Speed*, *Front-Side-Bus* and *L2-Cache*; for AMD processors, the interpretation is not as straightforward, but nonetheless it may be useful; and
  - *Size of Memory for Standalone Video Cards*.
- The websites from which data are currently being collected are quite good choices. However, other websites could be considered to ensure representativeness.

4.2 Fitting some theoretically appropriate models

Four regression models – henceforth referred to as 'Full (double-log)', 'Full (semi-log)', 'Streamlined (double-log)' and 'Streamlined (semi-log)' – were created as possible replacements for the hedonic regression model described in section 2.1 (henceforth referred to as 'Original'). As table 4.1 shows, the natural logarithm of price was modelled as a linear function of:

- All of the variables in 'Original', except for the interaction terms, the squared terms and *Software*. *Website* and an interaction between *Website* and *Vendor* were added, as was *Video Card Type* (where available);
- The above, with *RAM Size* and *Hard Drive Size* instead of the logarithmic versions;
- *Time Dummy*, *Vendor*, *Website*, an interaction between *Website* and *Vendor*, *CPU Type*, Log2(*RAM Size*), *Monitor Size* and (where available) *Video Card Type*; and
- The above, with *RAM Size* instead of the logarithmic version.

4.1 The regression models referred to in the paper

Regression model name	*Functional form*	*Explanatory variables*
Original	Double-log	*Time Dummy*, *Vendor*, *CPU Type*, *Log2(RAM Size)*, *Monitor Size*, *Log(Hard Drive Size)*, *CDRW*, *DVD*, *DVDRW*, *CDRW_DVD*, *Ethernet*, *Operating System*, *Software*, various interaction terms and squared terms.
Full (double-log)	Double-log	*Time Dummy*, *Vendor*, *Website*, *Vendor*Website*, *CPU Type*, *Log2(RAM Size)*, *Monitor Size*, *Log(Hard Drive Size)*, *CDRW*, *DVD*, *DVDRW*, *CDRW_DVD*, *Ethernet*, *Operating System*, *Video Card Type.*
Full (semi-log)	Semi-log	*Time Dummy*, *Vendor*, *Website*, *Vendor*Website*, *CPU Type*, *RAM Size*, *Monitor Size*, *Hard Drive Size*, *CDRW*, *DVD*, *DVDRW*, *CDRW_DVD*, *Ethernet*, *Operating System*, *Video Card Type*.
Streamlined (double-log)	Double-log	*Time Dummy*, *Vendor*, *Website*, *Vendor*Website*, *CPU Type*, *Log2(RAM Size)*, *Monitor Size*, *Video Card Type*.
Streamlined (semi-log)	Semi-log	*Time Dummy*, *Vendor*, *Website*, *Vendor*Website*, *CPU Type*, *RAM Size*, *Monitor Size*, *Video Card Type*.

Following comments from MAC (see the text boxes in Sections 2.3 and 3), some changes were made to the candidate models that were presented to MAC – specifically, weights were re-introduced, all important variables were forced in, and a bias correction was added. All of the output in this paper comes from these adjusted models.

4.3 Choosing the optimal model

All of the four candidate regression models are reasonable, given the information in Section 4.1. All include, as explanatory variables, the four attributes that the IT experts consulted considered to be the most important in explaining *Price*. Furthermore, none include forms of included explanatory variables believed to be inappropriate by the IT experts consulted, or forms shown to be inappropriate by exploratory data analysis (see Appendix B). The inclusion of an interaction between *CPU Type* and *RAM Size* was trialed following the advice of IT Client Service Section. Its effect on regression diagnostics was negligible, and thus it was not found to justify the increased complexity caused by its inclusion.

In addition, all of the four candidate regression models comply with economic theory which supports the use of double-log and semi-log functional forms in hedonic regressions. As discussed in Section 2.3, the double-log functional form in addition provides 'insurance' should large changes in computers' characteristics occur in the future. This may lead to a preference for the double-log candidate models.

All of the four candidate regression models exhibit good overall fit (based on inspections of plots of standardised residuals vs predicted values and leverage – see Appendix C for details) and no problematic multicollinearity (based on inspections of Variance Inflation Factors – see Appendix C for details). R^2 is, on average, 0.74 for the 'Full' models, 0.69 for 'Streamlined (double-log)', and 0.68 for 'Streamlined (semi-log)'. Adjusted-R^2 is, on average 0.73 for the 'Full' models, 0.68 for 'Streamlined (double-log)', and 0.67 for 'Streamlined (semi-log)'. Theory does not specify how large R^2 and Adjusted-R^2 must be in order for a regression model to be adequate, thus it is unclear whether these differences should lead one to prefer the 'Full' models.

The coefficients in all of the four candidate regression models are not always consistent with expected theoretical effects. As Section 2.3 explains, this is not necessarily alarming.

Minimising the number of explanatory variables increases ease of interpretation and simplicity. In addition, minimising the number of explanatory variables is extremely important from a cost point of view – reducing the number of variables collected reduces the time taken to collect the data, and reduces the time taken to train new staff to collect the data. In addition, reducing the number of variables collected allows, given fixed resources, data to be collected with greater accuracy and in more detail, and may allow the sample size to be increased. Thus, all other things being equal, one would prefer one of the 'Streamlined' models.

A regression model created to be an element of quarterly index processing must produce sensible estimates. Models which lead to the production of a hedonic index which moves as one would expect, given knowledge of the relevant market, are therefore preferred. All of the four candidate regression models lead to the production of such an index.

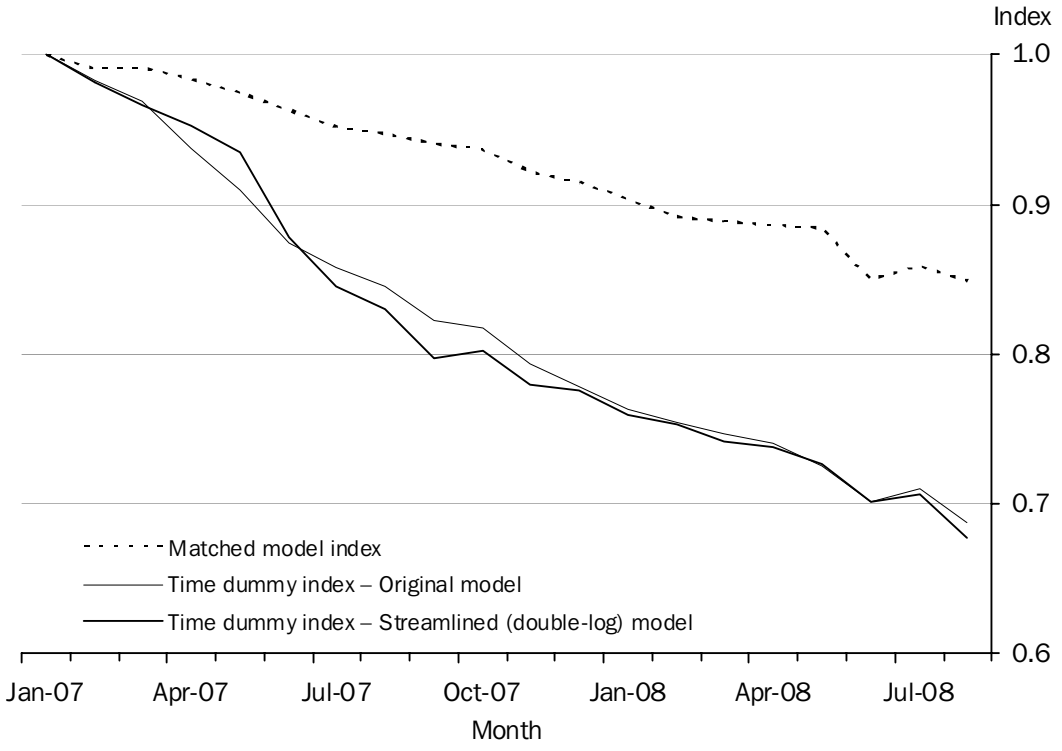
Figure 4.2 presents the matched model index and the time dummy indexes produced using the 'Original' and 'Streamlined (double-log)' models. (Appendix A explains how matched model indexes and time dummy indexes are calculated.) The corresponding graphs for the other candidate models (displayed in Appendix D.1) are similar. Given the highly competitive nature of the market for desktop computers, a set of expectations was conceived.

First, constant quality prices were expected to decrease over time. Accordingly, the time dummy index produced using the 'Streamlined (double-log)' model (as well as the time dummy index produced using the 'Original' model and the matched model index) shows a declining trend.

Second, constant quality prices of new and superseded models were expected to fall faster then than constant quality prices of matched models (NB this assumption has proved contentious). Accordingly, the time dummy index produced using the 'Streamlined (double-log)' model (as well as the time dummy index produced using the 'Original' model) declines faster than the matched model index.

Third, unusual changes (i.e. positive changes and large negative changes) in constant quality prices of new and superseded models were expected to occur only when such changes in constant quality prices of matched models occur. Accordingly, the time dummy index produced using the 'Streamlined (double-log)' model (as well as the time dummy index produced using the 'Original' model) only moves unusually when the matched model index does so.

4.2 Comparison of the matched model index and some time dummy indexes



In light of the above discussion, ASB has recommended that the 'Original' model be replaced with the 'Streamlined (double-log)' model. This change will, as explained in Appendix D.2, have very little impact on the published PPI APMI (ANZSIC 93, Group 284 – Electronic equipment manufacturing) series and the published CPI (Recreation group) series. In addition, this change will not represent a change in the fundamental methodology used by the ABS to calculate a price index for desktop computers (described in ABS, 2005).

5. EXTENDING THE RESULTS

One of the desired outcomes of the review of the desktop computer price index is the development of ideas regarding techniques which can be applied to the construction of hedonic price indexes for other consumer durables.

Due to technological change, the markets for all consumer durables change constantly. Thus, the hedonic regression models constructed for all consumer durables should be reviewed regularly. Regression models for consumer durables which change drastically with technological change (such as televisions) should be reviewed frequently, whilst regression models for more stable consumer durables (such as whitegoods) would not require review so often.

The process described in Section 3.1 could be used to perform the review. Persons with expert knowledge of the consumer durable in question would be consulted, candidate regression models would be fitted, then an 'optimal' model would be chosen, taking the following into consideration:

- the 'reasonableness' of the regression model, given industry knowledge;
- regression diagnostics;
- ease of interpretation and simplicity; and
- the 'reasonableness' of the index produced using the regression model.

In order to assess the 'reasonableness' of the index produced using the regression model, one must assert that the prices of unmatched models, of the good in question, can be expected to follow some pattern. In Section 4.3, a set of expectations, formed based on knowledge of the market for desktop computers, was put forward. These expectations cannot be extended to other consumer durables, sold in less competitive markets. However, it seems reasonable to assume that, unless something in the external environment upsets the balance, the relationship between price movements for matched models and price movements for unmatched models would remain constant. Thus, the 'reasonableness' of the index produced could be evaluated by monitoring whether the prices of unmatched models move unusually when the prices of matched models do not, or vice versa, and scanning the environment for an explanation when that occurred. Only when no explanation for the abnormality surfaced would one worry about the 'reasonableness' of the index.

6. FURTHER WORK

The Analytical Services Branch is currently using the knowledge built through the review of the desktop computer price index to build a de Haan double imputation hedonic price index for laptop computers. The Branch is also currently conducting further investigations into the issue of weighting. In addition, we intend to conduct further research into techniques which can be applied to the construction of hedonic price indexes for other consumer durables. The Producer Price Indexes Section is currently considering investigating data capture technology, to improve the quality of data as well as to reduce costs.

ACKNOWLEDGEMENTS

The authors would like to thank the following people for their contribution to this project: Ruel Abello, Wayne Qu, Shirley Clark, Susan Kluth, Leigh Merrington, Lewis Conn, Khann Moore, Keith Woolford, Laurie Nitschke, Darrol Dykes, Kevin Fox and Jan de Haan.

REFERENCES

Australian Bureau of Statistics (2005) *The Introduction of Hedonic Price Indexes for Personal Computers*, cat. no. 6458.0, ABS, Canberra.

Bascher, J. and Lacroix, T. (1999) "Dishwashers and PCs in the French CPI: Hedonic Modeling, from Design to Practice", Paper presented at the Fifth Meeting of the International Working Group on Price Indices, Reykjavik 25–27 August.

de Haan, J. (2003) "Time Dummy Approaches to Hedonic Price Measurement", Paper presented at the Seventh Meeting of the International Working Group on Price Indexes (Ottawa Group), Paris 27–29 May.

Diewert, E. (2003) "Hedonic Regressions: A Review of Some Unresolved Issues", Paper presented at the Seventh Meeting of the International Working Group on Price Indexes (Ottawa Group), Paris 27–29 May.

Diewert, E.; Heravi, S. and Silver, M. (2007) "Hedonic Imputation versus Time Dummy Hedonic Indexes", University of British Columbia Discussion Paper.

International Monetary Fund (2004) *Producer Price Index Manual*, IMF, Washington, D.C..

Lim, P. and McKenzie, R. (2001) "Hedonic Price Analysis for Personal Computers in Australia: An Alternative Approach to Quality Adjustments in the Australian Price Indexes", Internal paper, Australian Bureau of Statistics, Canberra.

Pakes, A. (2003) "A Reconsideration of Hedonic Price Indexes with Application to PCs", *American Economic Review*, 93(5), pp. 1578–1596.

Purcell, J.; Esguerra, E. and Branson, M. (2007) "Improving the Hedonic Model for Personal Computers (for use in the CPI and PPI)", Internal paper, Australian Bureau of Statistics, Canberra.

Silver, M. (2003) "The Use of Weights in Hedonic Regressions: The Measurement of Quality-Adjusted Price Changes", Paper presented at the Seventh Meeting of the International Working Group on Price Indexes (Ottawa Group), Paris 27–29 May.

Triplett, J. (2004) "Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products", OECD STI Working Paper 2004/9.

Turnbull, J. (2006) "Background – Review of Hedonic Computer Price Index", Internal paper, Australian Bureau of Statistics, Canberra.

GLOSSARY

This paper uses the following terminology, believed to be the least confusing, though we note that it is not universally applied throughout the literature.

Term	*Definition*
Attribute	An element of a good, through which the good satisfies the needs or wants of households or the community or produces other goods or services.
De Haan double imputation hedonic price index	A price index wherein price changes for matched models are calculated using the matched model index approach, whilst price changes for unmatched models calculated using the time dummy index approach.
Double-log regression model	A regression model wherein the Response variable is a logarithmic form of the variable of interest, and the continuous Explanatory variables are also in logarithmic form.
Good	A physical object for which a demand exists, over which ownership rights can be established, and whose ownership can be transferred from one institutional unit to another by engaging in transactions on the market. Goods are in demand because they may be used to satisfy the needs or wants of households or the community or used to produce other goods or services (IMF, 2004). Many goods (such as desktop computers) are sold in the form of models, characterised by attributes.
Hedonic function	A function (often a regression model) which relates the price of a good to its attributes.
Hedonic price index	A price index that utilises, in some manner, a hedonic function.
Matched model	A model which exists in both of the two time periods in question.
Matched model index	A 'geometrically weighted two period matched model index' - a price index calculated from price changes observed for matched models.
Model	A variety of a good, with some fixed set of attributes, often described by a 'model number'. Not to be confused with regression model.
Regression model	A linear regression model – an equation which expresses the expected value of the response variable as a linear function of some explanatory variables, estimated using the 'weighted least squares' method.
Semi-log regression model	A regression model wherein the response variable is a logarithmic form of the variable of interest, and the continuous explanatory variables are in untransformed form.
Time dummy index	A 'two-period pooled time dummy index' – a price index calculated from price changes imputed from the coefficient of the 'time dummy' explanatory variable in a two period pooled hedonic regression model.
Unmatched model	A model which exists in only one of the two time periods in question. Unmatched models are also referred to as 'new and superseded models' or 'births and deaths'.

APPENDIXES

A. DOUBLE IMPUTATION HEDONIC APPROACH TO INDEX CONSTRUCTION

The ABS introduced a de Haan double imputation hedonic price index for desktop computers into the Producer Price Indexes in 2003 (ABS, 2005) and into the Consumer Price Index and the National Accounts in 2005 (Purcell, Esguerra and Branson, 2007). Under the de Haan double imputation approach, price changes for matched models are calculated from observed price changes, whilst price changes for unmatched models (i.e. superseded and new models) are imputed from a hedonic function. This Appendix briefly explains how this is done.

A.1 Some nomenclature and notation

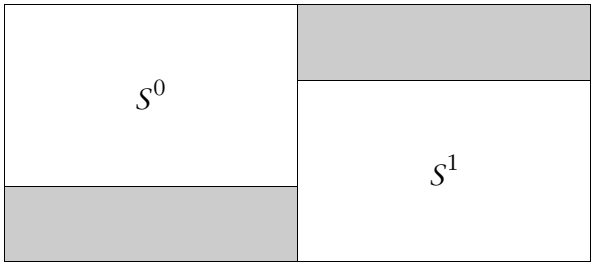
For each time period t, we consider a sample of models. For each model i we observe:

a period t price p $\quad p_i^t$

and a set of k attributes z: $\quad \underset{\sim}{z}_i^t = \left(z_{i1}, \ldots, z_{ik}\right)$

We denote this set S by $\quad S^t = \{p_i^t, \underset{\sim}{z}_i^t\}$

Say we now consider two time periods, 0 and 1. We denote this diagrammatically by



Some models are in S^0 and S^1 – we refer to these as matched models. Associated with each of these models is a set of attributes and two prices (a price from period 0 and a price from period 1).

We consider here two different overlapping sets, M^0, the set of matched models in period 0 and M^1, the set of matched models in period 1:

$$M^0 = \{p_i^0, z_i^0, i \in S^0 \cap S^1\}$$

$$M^1 = \{p_i^1, z_i^1, i \in S^0 \cap S^1\}$$

Some models are in S^0 only – we refer to these as superseded models, or deaths. We consider here the set of deaths D^0 at time 0:

$$D^0 = \{p_i^0, z_i^0, i \in S^0 \not\subset S^1\}$$

Some models are in S^1 only – we refer to these as new models or births. We consider here the set of births B^1 at time 1:

$$B^1 = \{p_i^1, z_i^1, i \in S^1 \not\subset S^0\}$$

The relationships between the above sets can be described in the following diagram:

S^0	S^1
D^0	
M^0	M^1
	B^1

Models are sampled from n vendors. Denote the weights assigned to those vendors by w_k, $k = \{1, \ldots, n\}$.

Denote the number of sampled observations (of each vendor k) at $t = 1$ by n_{1k}. Then the unit weight of each observation i (of vendor k) in time $t = 1$ is

$$w_{1t} = \frac{w_k}{n_{1k}}.$$

A.2 Calculating price changes for matched models

Consider the matched models, $i \in S^0 \cap S^1$.

For each matched model i, we calculate a ratio r

$$r_i = \frac{p_i^1}{p_i^0}$$

Next, for each vendor k we calculate quantities X and Y

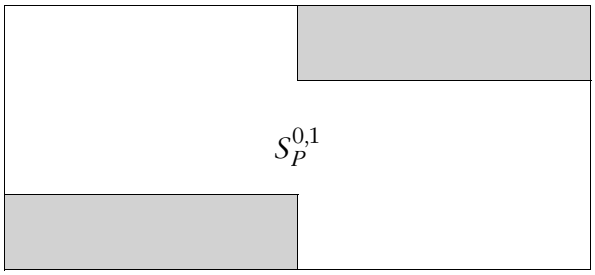
$$X_k = \sum_{i \in M^1} \log r_i \cdot w_{i1} \cdot I_{i \text{ is vendor } k}$$
$$Y_k = \sum_{i \in M^1} w_{i1} \cdot I_{i \text{ is vendor } k}$$

Finally, we calculate the geometrically weighted two-period matched model index value for $t = 1$ as

$$P_M^1 = \exp\left(\frac{X_k}{Y_k}\right) \cdot 100$$

A.3 Imputing price changes for unmatched models

Consider the union of S^0 and S^1. Denote this by $S_P^{0,1}$.



A hedonic function is estimated by performing regression on $S_P^{0,1}$.

The regression is performed using weighted least squares, with w_k, $k = \{1, 2, 3\}$, used for weighting.

Consider the double-log regression model

$$\log p_i^t = \beta_0 + \delta d_i + \sum_{j=1}^{m} \beta_j \log z_{ji} + \sum_{j=m+1}^{k} \beta_j z_{ji} + \varepsilon_i$$

where $$d_i = \begin{cases} 1 & i \in S^1 \\ 0 & \text{otherwise} \end{cases}$$

is the 'time dummy';

$$\left(z_{i1},\ldots,z_{im}\right)$$

are the included continuous variables; and

$$\left(z_{im+1},\ldots,z_{ik}\right)$$

are the included dummy variables.

Also, consider the semi-log regression model

$$\log p_i^t = \beta_0 + \delta d_i + \sum_{j=1}^{k} \beta_j z_{ji} + \varepsilon_i$$

In either case, $$\log \frac{\hat{p}^1}{\hat{p}^0} = \hat{\delta}$$

Thus, we can calculate a time dummy index value for $t = 1$ as

$$P_{TD}^1 = \exp \hat{\delta} \cdot 100$$

The time dummy index produced by the 'Original' model is calculated in this way.

Examination of the double-log and semi-log regression models shows that they produce unbiased estimates of log(price), but upwards biased estimates of price. One expects the bias to be small, but should nonetheless correct for it as follows.

As $\hat{\delta}$ has a normal distribution, $\exp\hat{\delta}$ has a lognormal distribution and so

$$E\left(\exp\hat{\delta}\right) = \exp\left(\delta + \tfrac{1}{2}\operatorname{var}\hat{\delta}\right).$$

Thus, $\exp\delta$ should be estimated as

$$\exp\left(\hat{\delta} - \tfrac{1}{2}\widehat{\operatorname{var}}\hat{\delta}\right)$$

A time dummy index value for time $t = 1$ can then be calculated as

$$P_{TD}^{1} = \exp\left(\hat{\delta} - \tfrac{1}{2}\widehat{\text{var}}\hat{\delta}\right) \cdot 100$$

The time dummy indexes produced by the 'Full (double-log)', 'Full (semi-log)', 'Streamlined (double-log)' and 'Streamlined (semi-log)' models are calculated in this way.

A.4 Calculating the de Haan double imputation index

We calculate f_M, the fraction matched, as

$$f_M = \frac{\left|M^1\right|}{\left|S^1\right|}$$

Following de Haan (2003), we calculate the de Haan double imputation hedonic price index value for $t = 1$ as

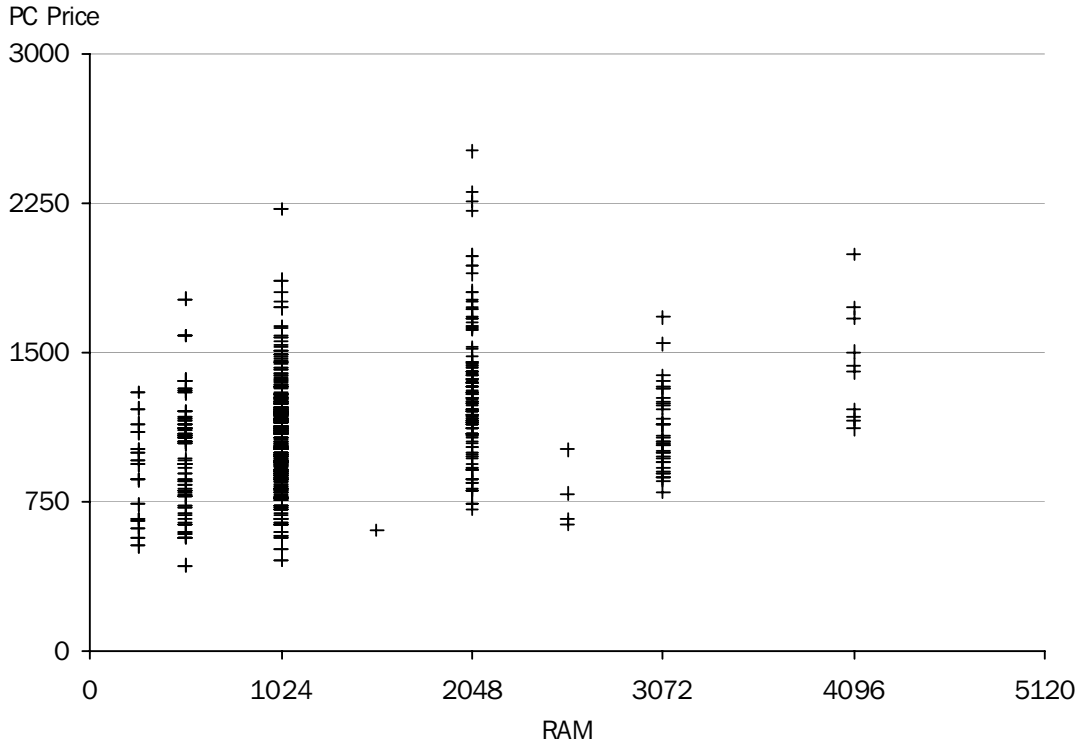
$$P_{DI}^{1} = \left[P_M^1\right]^{f_M} \cdot \left[P_{TD}^{1}\right]^{1-f_M}.$$

B. EXPLORATORY DATA ANALYSIS

This Appendix presents exploratory data analysis output for May/June 2008. This output is representative of the output for all periods considered.

Plots of *Price* vs collected characteristics were constructed to explore relationships. However, as most computer attributes are discrete in nature, it is difficult to see clear relationships, as one can see in figure B.1.

B.1 Plot of *Price* vs *RAM*



C. REGRESSION OUTPUT

This Appendix presents regression output for May/June 2008, for the 'Streamlined (double-log)' model. This output is representative of the output for all periods considered, for all candidate models.

C.1 Diagnostics

Diagnostic plots are presented in figures C.1 and C.2. Figure C.1 presents a plot of standardised residuals vs predicted values. It shows that there are no trends in the residuals – this indicates that the functional form is appropriate. It also shows that there is no heteroscedasticity – this indicates that the regression assumption of constant variance is not violated.

C.1 Standardised residuals vs predicted values

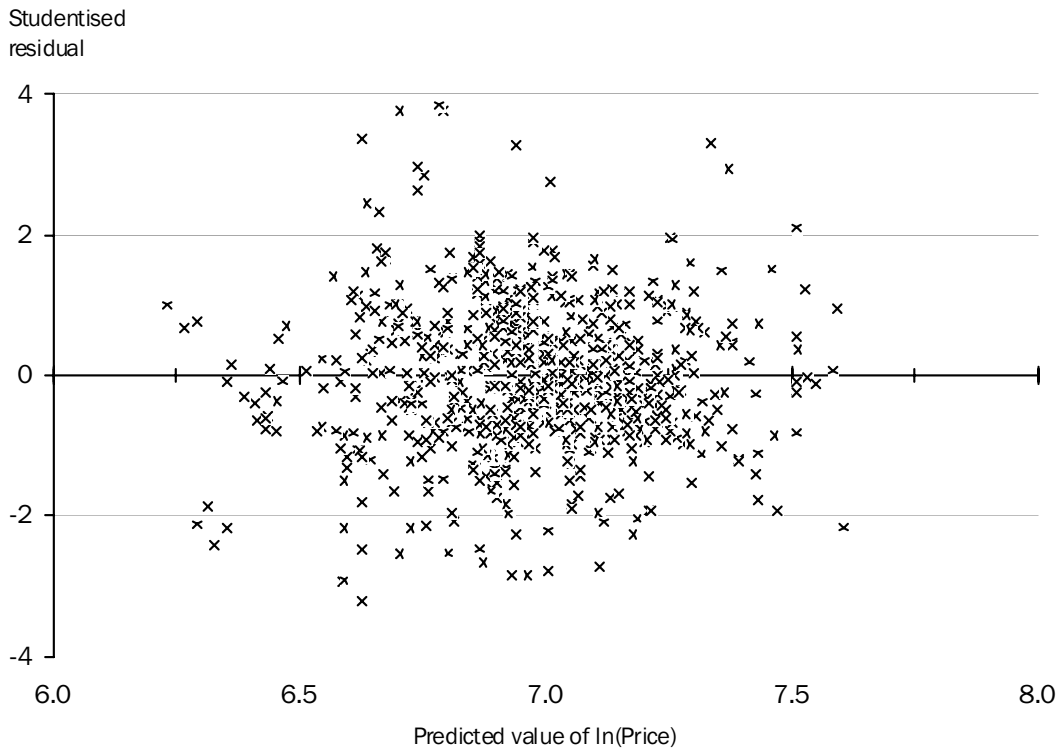
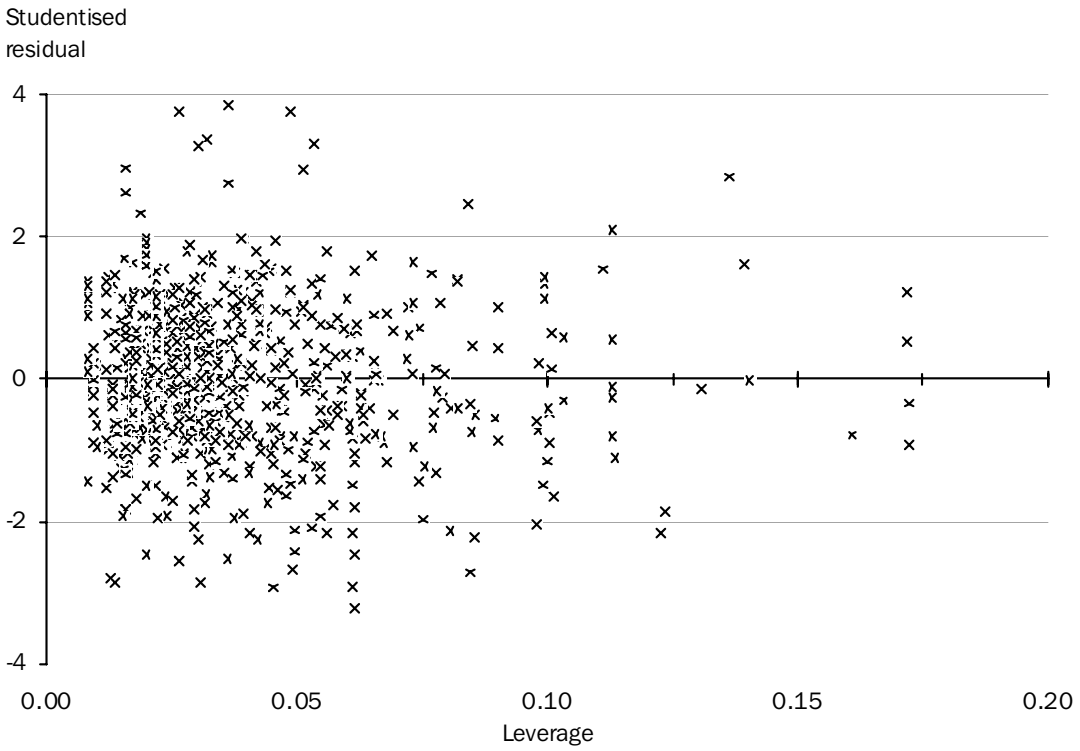


Figure C.2 presents a plot of standardised residuals vs leverage. It shows that there are no points with a large leverage and a large absolute residual. As Silver (2003) explains, this indicates that the model is not unduly skewed by any observation.

C.2 Standardised residuals vs leverage



R^2 is 0.7195 and Adjusted-R^2 is 0.7095. Theory does not specify how large these values must be, however the authors believe that they are sufficiently large.

C.2 Parameter estimates

Table C.3 presents parameter estimates. It shows that no parameter estimates have Variance Inflation Factors much greater than the rule-of-thumb cut-off of 10, and thus shows that there is no problematic multicollinearity.

C.3 Parameter estimates for the 'Streamlined (double-log)' model, May/June 2008

Variable	*Parameter estimate*	*Standard error*	*Pr > |t|*	*Variance inflation*
Intercept	5.9989	0.1069	<0.0001	0.0000
timedummy	–0.0358	0.0100	0.0003	1.0561
Vendor2	0.0854	0.0164	<0.0001	2.1555
Vendor3	–0.1221	0.0280	<0.0001	5.1197
Website2	0.0892	0.0229	0.0001	3.4564
Website3	0.3004	0.0363	<0.0001	6.2470
Vendor2* Website2	–0.0764	0.0345	0.0273	2.0216
Vendor3* Website2	–0.0268	0.0376	0.4762	4.3163
Vendor3* Website3	0.1377	0.0413	0.0009	3.8746
CPU2	–0.1882	0.0378	<0.0001	2.4145
CPU3	–0.0714	0.0333	0.0326	3.7256
CPU4	–0.0828	0.0483	0.0868	1.5805
CPU5	–0.3450	0.0279	<0.0001	3.0928
CPU6	0.0672	0.0295	0.0227	9.2275
CPU7	0.3055	0.0407	<0.0001	2.2460
CPU8	–0.1065	0.0324	0.0011	4.4638
log2ram	0.0780	0.0093	<0.0001	2.5532
monitor2	0.2878	0.0540	<0.0001	1.2121
monitor3	0.2064	0.0169	<0.0001	1.2624
monitor4	0.2456	0.0136	<0.0001	1.3150
monitor5	0.2698	0.0222	<0.0001	1.2314
monitor6	0.3389	0.0221	<0.0001	1.2629
monitor7	0.6699	0.0573	<0.0001	1.1274
graphics2	–0.0216	0.0478	0.6518	1.9826
graphics3	0.0146	0.0383	0.7033	3.2580
graphics4	–0.0063	0.0377	0.8673	3.8164
graphics5	0.0058	0.0351	0.8682	4.5242
graphics6	0.0327	0.0315	0.2987	10.5427
graphics7	0.0160	0.0383	0.6759	3.1659
graphics8	0.0550	0.0349	0.1154	5.5016
graphics9	0.0672	0.0383	0.0794	3.0409

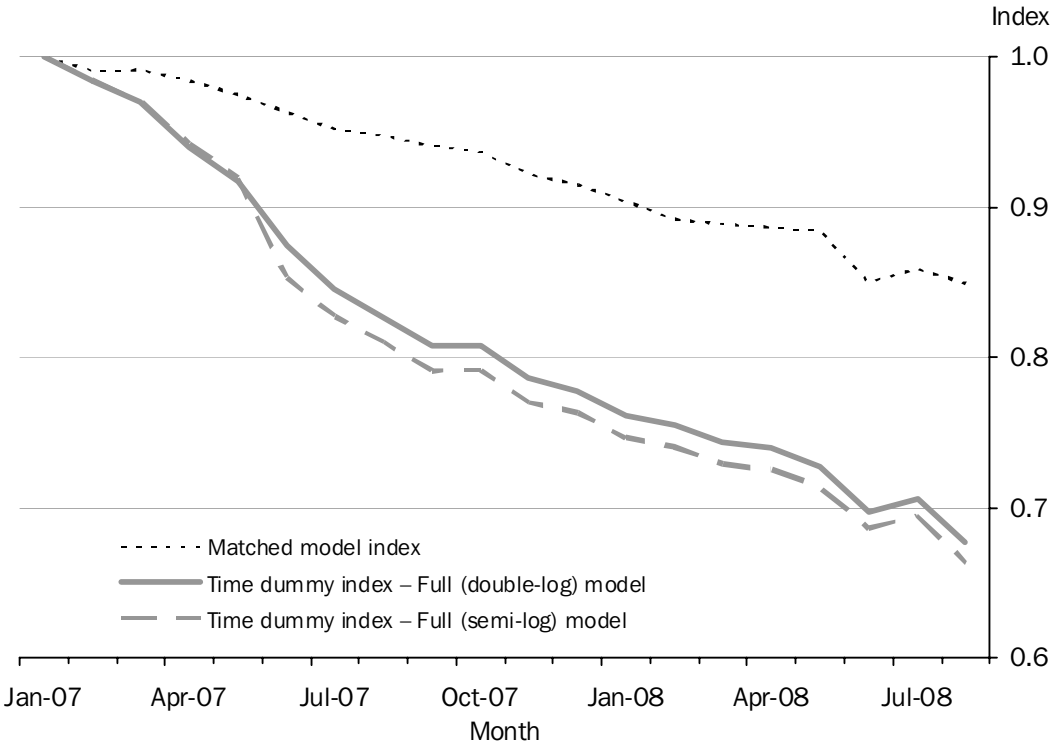
D. COMPARING THE INDEXES PRODUCED

This Appendix presents some investigations of the 'reasonableness' of the indexes produced using the four candidate models.

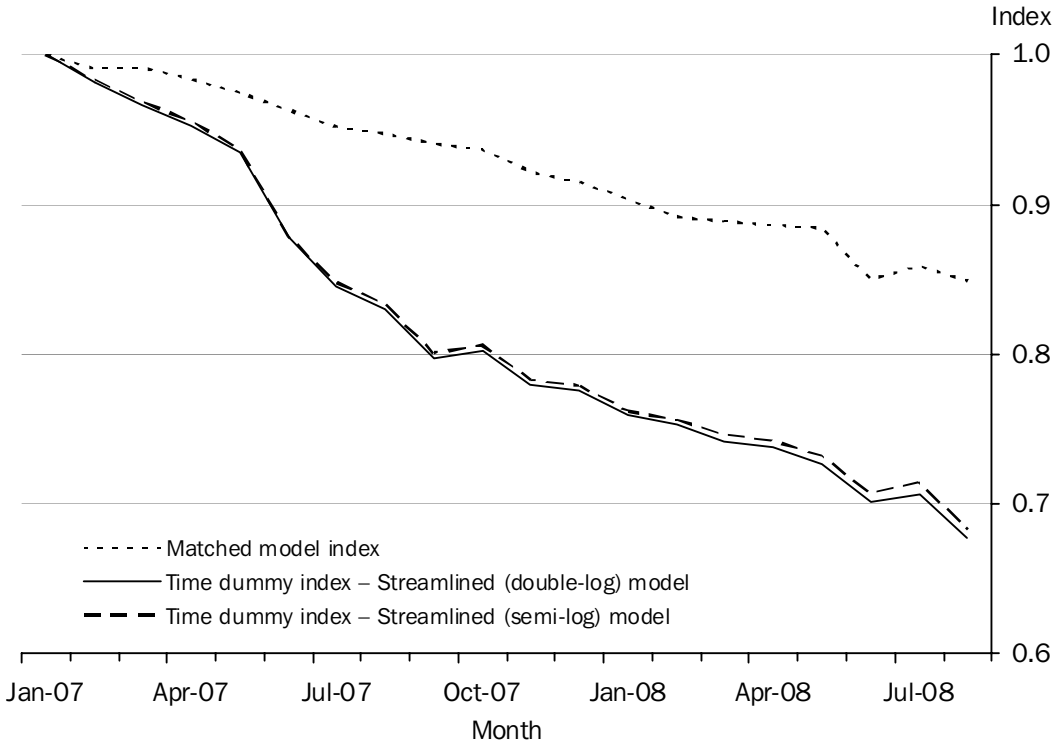
D.1 Comparing the time dummy indexes produced

The time dummy indexes produced by the four candidate models and the original model are compared in this section. Figures D.1 and D.2 compare the time dummy indexes produced by semi-log and double-log versions of the 'Full' and 'Streamlined' models respectively. They show that the differences between the indexes produced by the two versions are very minor, particularly for the 'Streamlined' model.

D.1 Comparison of Full (double-log) and Full (semi-log) time dummy indexes



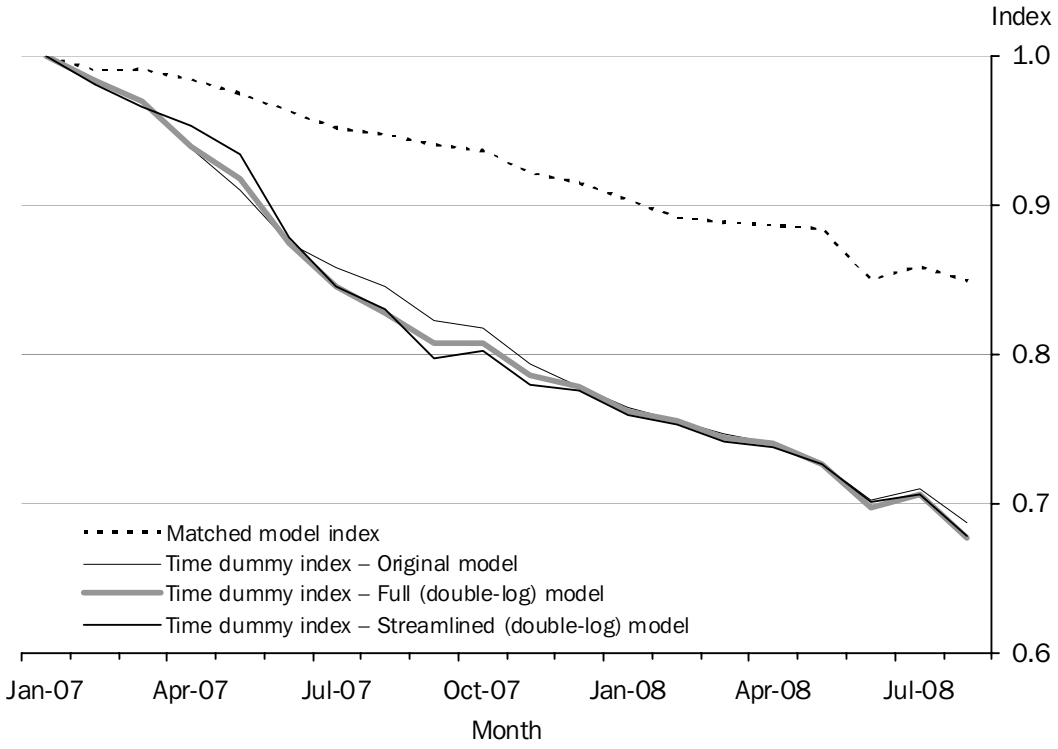
D.2 Comparison of Streamlined (double-log) and Streamlined (semi-log) time dummy indexes



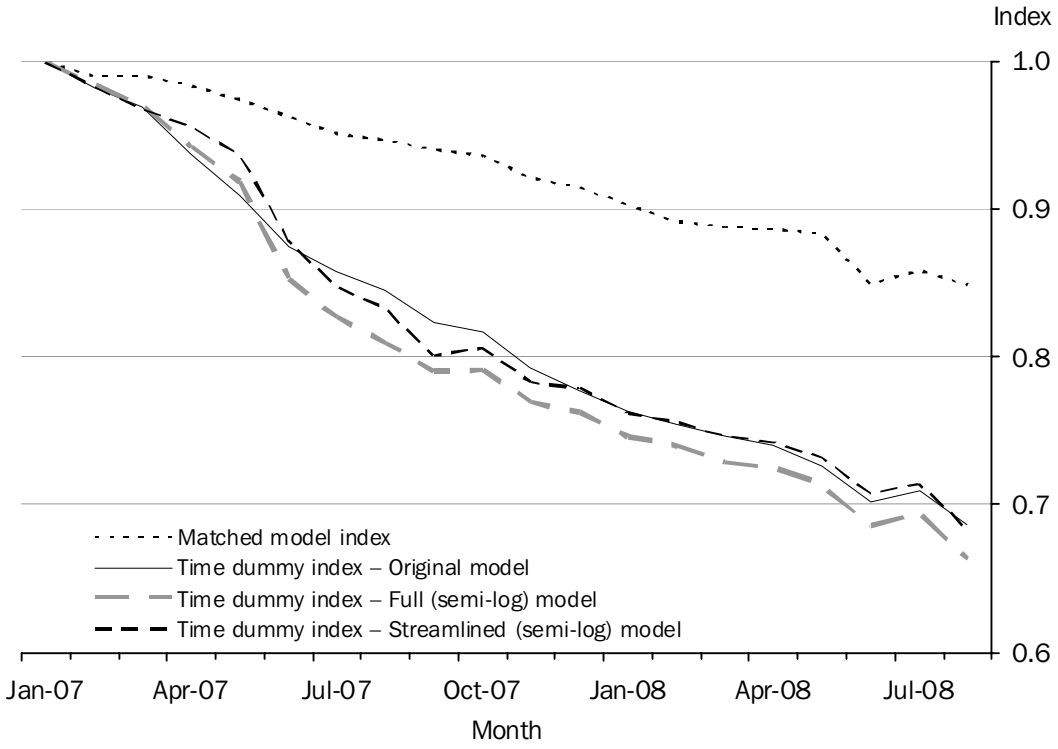
Figures D.3 and D.4 compare the time dummy indexes produced by the 'Full' and 'Streamlined' versions of the double-log and semi-log models respectively, with the time dummy index produced by the 'Original' model. They show that the differences between the time dummy indexes produced by all of these models are relatively minor.

Section 4.3 explained why the time dummy index produced by 'Streamlined (double-log)' is reasonable. As the time dummy indexes produced by all of the candidate models have few differences, all of the candidate models similarly meet the 'reasonable index' criterion.

D.3 Comparison of Full (double-log) and Streamlined (double-log) time dummy indexes



D.4 Comparison of Full (semi-log) and Streamlined (semi-log) time dummy indexes



D.2 Comparing the double imputation indexes produced

The double imputation index produced using the 'Streamlined (double-log)' model is compared with that produced by the 'Original' model in this section. The double imputation indexes produced by the other candidate models are not shown, but are similar.

As Appendix A explains, a de Haan double imputation index is a geometric mean of a matched model index and a time dummy index, weighted by the fraction of observations matched between the two periods. The choice of hedonic model thus affects the de Haan double imputation index only through its effect on imputed price changes for unmatched models – its effect is greater when the percentage of sampled models which are matched models is less. Figure D.5 shows the proportion of matched models in each month's sample. The average proportion is 0.69.

D.5 Percentage of matched models in the monthly sample

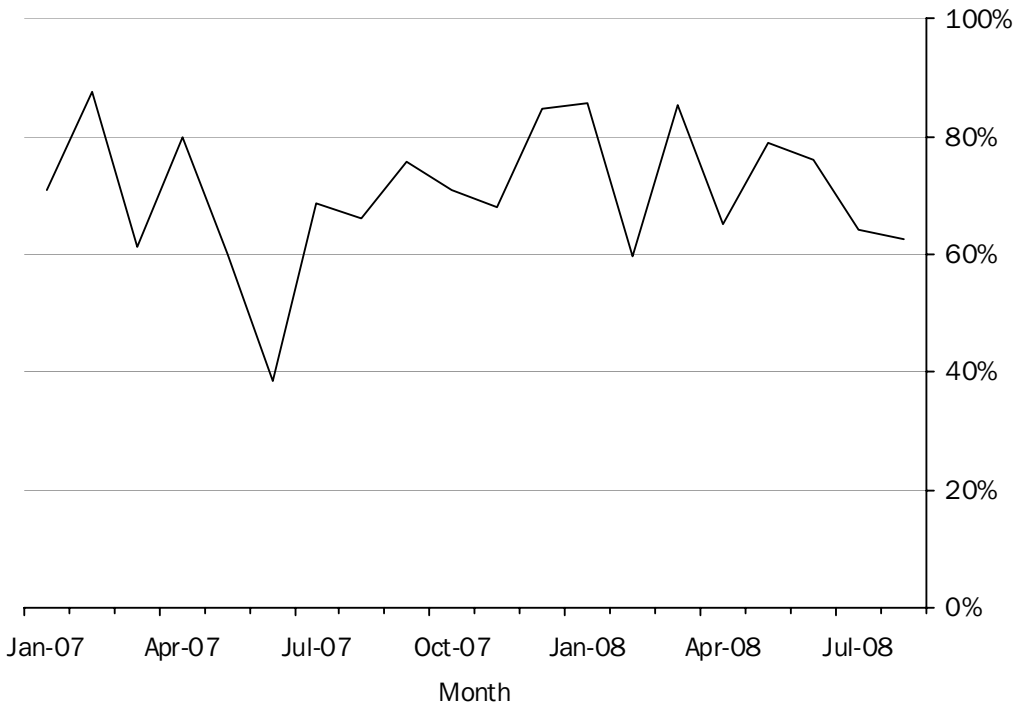
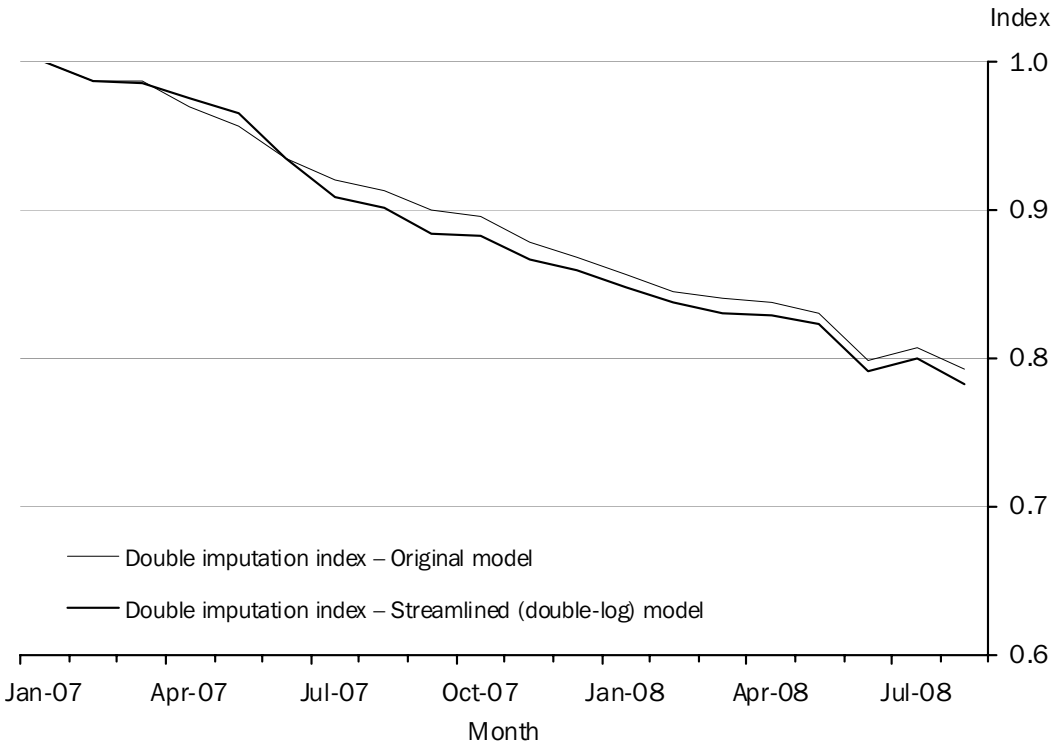


Figure D.6 presents the de Haan double imputation indexes produced using the 'Original' and 'Streamlined (double-log)' models. It shows these indexes are very similar. Thus, switching from 'Original' to 'Streamlined (double-log)' should have very little impact on the published PPI APMI (ANZSIC 93, Group 284 – Electronic equipment manufacturing) series and the published CPI (Recreation group) series. Note, however, that the index numbers actually used to construct the aforementioned published series may differ from those shown in figure D.6, due to editing.

D.6 Comparison of Original and Streamlined (double-log) double imputation indexes



FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au