THE "HEDONIC" MODEL USED IN THE ODPM'S HOUSE PRICE INDICES

A description of the regression model developed for the new ODPM House Price Index that was launched in September 2003

The new monthly House Price Index published by ODPM is based on a multivariate fixed effects regression model ("hedonic" model) of house prices. This paper provides a technical description of the underlying model and how it may be used to estimate average house prices, house price movements and the house price index. For clarity of exposition, the paper contains the following sections:

1. Introduction
2. Specification of the Hedonic Regression Model
3. Estimation of the Expected House Price for a Cell
4. Estimation of the House Price Index
5. Weighted Regressions

Appendix - Estimation of Cumulants

1. Introduction

The rationale for using the hedonic regression approach is that the number of "cells" (that is, possible combinations of different sets of values for the main effects considered) is so large that it is not possible to estimate accurately, or at all, average house prices for each cell individually from the data available. We assume that the influence of each main effect or interaction is the same across all cells and we include in the model those main effects and interactions which "best" explain variation in house prices. The concept of "best" is determined according to Schwarz's Bayesian Criterion (see Schwarz, 1978 and Shi & Tsai, 2002).

The model used includes the following seven main effects:

- location (local authority district or London borough)
- local authority cluster (an ONS classification of local authorities)
- type of neighbourhood (ACORN)
- dwelling type
- number of habitable rooms (or bedrooms)
- old/new
- first-time buyer/former owner occupier (FTB/FOO)

and the following three interactions:

- ACORN × dwelling type
- ACORN × FTB/FOO
- dwelling type × old/new.

For ease of notation, we denote each cell by a single subscript *i* (or *j*). Associated with each cell *i* is a vector $\boldsymbol{x}_{yi}$ containing the values of the variables in the regression model for cell *i*. The subscript *y* indicates the calendar year to which the data relate. This is because the new house price index is annually chain-linked and we need to allow for the possibility that the model may change at the January link months (because more housing characteristics become available, for example).

Note that, because all the main effects in the model are categorical, they are represented by dummy, dichotomous covariates and the vector $\boldsymbol{x}_{yi}$, contains only the values 1 or 0, according to whether the relevant characteristic is present or not in the cell. Note also that some elements of $\boldsymbol{x}_{yi}$ are mutually associated (for example, a dwelling type cannot be both detached and terraced) and all covariates must be linearly independent, to ensure the validity of matrix inversion in estimation. Thus the main effect old/new will be represented by a single covariate with the value (for example) 0 for old and 1 for new.

2. Specification of the Hedonic Regression Model

In month *m* of year *y*, we assume the following model for the price P_{ymik} of dwelling *k* in cell *i*:

$$L_{ymik} = \ln\left(P_{ymik}\right) = \boldsymbol{x}'_{yi}\boldsymbol{\beta}_{ym} + \varepsilon_{ymik} \qquad 2.1$$

where: $\boldsymbol{\beta}_{ym}$ is a column vector of p_y unknown parameters which determine the proportionate impact on the expected price of the p_y dichotomous covariates included in the column vector $\boldsymbol{x}_{yi}$;

and the ε_{ymik} are independent, identically distributed random variables with the following properties:

$$E\left[\varepsilon_{ymik}\right] = 0; \quad Var\left[\varepsilon_{ymik}\right] = \sigma^2_{ym}; \quad Cov\left[\varepsilon_{ymik}, \varepsilon_{znjl}\right] = 0, \text{ for } ymik \neq znjl \qquad 2.2$$

We also assume that all the higher order central moments of ε_{ymik} exist. This is a necessary, but plausible, assumption to cater for the observed circumstance that the ε_{ymik} are not Normally distributed. In particular, we assume:

$$E\left[\varepsilon^3_{ymik}\right] = \mu_{3\,ym} = \kappa_{3\,ym}; \quad E\left[\varepsilon^4_{ymik}\right] = \mu_{4\,ym} = \kappa_{4\,ym} + 3\sigma^4_{ym}$$

The expected price of a dwelling in cell *i*, in month *m* of year *y* is then:

$$P_{ymi} = E\left[P_{ymik}\right] = E\left[\exp\left(L_{ymik}\right)\right] = \exp\left(\boldsymbol{x}'_{yi}\boldsymbol{\beta}_{ym}\right) E\left[\exp\left(\varepsilon_{ymik}\right)\right]$$

From the standard theory of moment generating functions (see, for example, chapter 3 of Stuart & Ord, 1987), we have, for random variable *x* and auxiliary variable *t*:

$$\ln\left(E\left[\exp\{xt\}\right]\right) = \sum_{r=1}^{\infty} \frac{\kappa_r t^r}{r!} \quad \text{That is: } E\left[\exp\{xt\}\right] = \exp\left(\sum_{r=1}^{\infty} \frac{\kappa_r t^r}{r!}\right) \qquad 2.3$$

where κ_r is the r^{th} cumulant of the distribution of *x*.

For the house price data and models examined, $\kappa_2 = \sigma^2 \approx 0.15$, $\kappa_3 \approx 0$, $\kappa_4 \approx 0.03$. Higher order cumulants are negligible and may be ignored.

For $x=\varepsilon_{ymik}$ and $t=1$, we therefore have:

$$E\left[\exp\left(\varepsilon_{ymik}\right)\right] \approx \exp\left(0+\frac{\sigma_{ym}^2}{2!}+\frac{\kappa_{3ym}}{3!}+\frac{\kappa_{4ym}}{4!}\right)$$

$$= \exp\left(\frac{\gamma_{ym}\sigma_{ym}^2}{2}\right)$$

where $\gamma_{ym} = 1+\frac{\kappa_{3ym}}{3\sigma_{ym}^2}+\frac{\kappa_{4ym}}{12\sigma_{ym}^2}$ 2.4

is an adjustment factor to correct for non-Normality.

Thus: $P_{ymi} = E\left[P_{ymik}\right] = \exp\left(\boldsymbol{x}'_{yi}\boldsymbol{\beta}_{ym}+\frac{1}{2}\gamma_{ym}\sigma_{ym}^2\right)$ 2.5

3. Estimation of the Expected House Price for a Cell

In practice, the vector $\boldsymbol{\beta}_{ym}$ is not known and has to be estimated. For this purpose, we use the following notation:

$\boldsymbol{l}_{ym}$= column vector of all the n_{ym} elements L_{ymik} in month m of year y;

$\boldsymbol{X}_{ym}$= matrix of housing characteristics for the sample in month m of year y: each row of the matrix is the row vector $\boldsymbol{x}'_{yi}$ of housing characteristics for the corresponding element in vector $\boldsymbol{l}_{ym}$;

$\boldsymbol{\varepsilon}_{ym}$= column vector of all the n_{ym} random components ε_{ymik} in month m of year y.

Equation 2.1 above relates to a single dwelling. The n_{ym} equations for month m of year y may be summarised in the matrix equation: $\boldsymbol{l}_{ym} = \boldsymbol{X}_{ym}\boldsymbol{\beta}_{ym} + \boldsymbol{\varepsilon}_{ym}$

with: $E\left[\boldsymbol{\varepsilon}_{ym}\right] = \boldsymbol{0}; \quad Var\left[\boldsymbol{\varepsilon}_{ym}\right] = \sigma^2_{ym}\boldsymbol{I}$ 3.1

Using ordinary least squares estimation, the estimator for $\boldsymbol{\beta}_{ym}$ is:

$$\hat{\boldsymbol{\beta}}_{ym} = \left(\boldsymbol{X}'_{ym}\boldsymbol{X}_{ym}\right)^{-1}\boldsymbol{X}'_{ym}\boldsymbol{l}_{ym} \qquad 3.2$$

with: $Var\left(\hat{\boldsymbol{\beta}}_{ym}\right) = \sigma^2_{ym}\left(\boldsymbol{X}'_{ym}\boldsymbol{X}_{ym}\right)^{-1} = \boldsymbol{V}_{ym}$ (for notational convenience) 3.3

(see, for example, Johnston, 1972).

In practice, the regression model for the House Price Index uses weighted least squares estimation in order to cater for missing values for some covariates. Section 5 below describes the reasons for this, the treatment applied, the impact on the model and amended formulae for $\hat{\boldsymbol{\beta}}_{ym}$ and $Var\left(\hat{\boldsymbol{\beta}}_{ym}\right)$.

We may therefore consider the following estimator for P_{ymi} :

$$\tilde{P}_{ymi} = \exp\left(\boldsymbol{x}'_{yi}\hat{\boldsymbol{\beta}}_{ym} + \tfrac{1}{2}\gamma_{ym}\sigma^2_{ym}\right) \qquad 3.4$$

where the unknown vector $\boldsymbol{\beta}_{ym}$ is replaced by the estimator $\hat{\boldsymbol{\beta}}_{ym}$ from equation 3.2.

However, because the vector $\hat{\boldsymbol{\beta}}_{ym}$ is random, the expected value of $\tilde{P}_{ymi}$ is obtained according to equation 2.3 above. We assume that the scalar random variable $\boldsymbol{x}'_{yi}\hat{\boldsymbol{\beta}}_{ym}$ is approximately Normal, because the number of parameters in $\hat{\boldsymbol{\beta}}_{ym}$ is very much

less than the number of observations on which the estimates are based. With this approximation, only the first two cumulants are non-zero and we have:

$$E\left[\tilde{P}_{ymi}\right] = \exp\left(\boldsymbol{x}'_{yi}\boldsymbol{\beta}_{ym} + \tfrac{1}{2}Var\left[\boldsymbol{x}'_{yi}\hat{\boldsymbol{\beta}}_{ym}\right] + \tfrac{1}{2}\gamma_{ym}\sigma^2_{ym}\right)$$

$$= \exp\left(\boldsymbol{x}'_{yi}\boldsymbol{\beta}_{ym} + \tfrac{1}{2}\boldsymbol{x}'_{yi}\boldsymbol{V}_{ym}\boldsymbol{x}_{yi} + \tfrac{1}{2}\gamma_{ym}\sigma^2_{ym}\right) \qquad 3.5$$

The estimator $\tilde{P}_{ymi}$ is therefore biased upwards by the factor $\exp\left(\tfrac{1}{2}\boldsymbol{x}'_{yi}\boldsymbol{V}_{ym}\boldsymbol{x}_{yi}\right)$. An approximately unbiased estimator is therefore:

$$\hat{P}_{ymi} = \exp\left(\boldsymbol{x}'_{yi}\hat{\boldsymbol{\beta}}_{ym} - \tfrac{1}{2}\boldsymbol{x}'_{yi}\boldsymbol{V}_{ym}\boldsymbol{x}_{yi} + \tfrac{1}{2}\gamma_{ym}\sigma^2_{ym}\right) \qquad 3.6$$

To estimate σ^2_{ym}, γ_{ym} and $\boldsymbol{V}_{ym}$, we use sample estimates of σ^2_{ym}, κ_{3ym} and κ_{4ym}. Because these variance and higher order cumulant terms are also estimated, additional bias correction terms should also, in principle, be included. However, these additional bias correction terms are of the order κ_{4ym}/n_{ym} or less. In practice, κ_{4ym} itself is small and, since n_{ym} is about 20,000, the additional bias correction term is negligible and may be ignored. Usually, $\hat{\gamma}_{ym}$ is of the order 1.01.

The appendix to this paper provides formulae, and their derivations, for estimating κ_{3ym} and κ_{4ym}.

4. Estimation of the House Price Index

To calculate the house price index, we calculate for each month a weighted average price for the domain of interest, where the domain may be all dwellings covered by the model or a subset of these dwellings, such as dwellings within a region, old or new dwellings, dwellings of a specific type. We use the estimated mean price in each cell from equation 3.6 and a set of weights $\{w_{Dyi}\}$, fixed for each calendar year and with $\sum_{i\in D} w_{Dyi} = 1$, where D denotes the set of cells included within the domain of interest. The weighted average price for domain D in month m of year y is then:

$$\bar{P}_{Dym} = \sum_{i\in D} w_{Dyi}\hat{P}_{ymi} = \sum_{i\in D} w_{Dyi}\exp\left(\boldsymbol{x}'_{yi}\hat{\boldsymbol{\beta}}_{ym} - \tfrac{1}{2}\boldsymbol{x}'_{yi}\hat{\boldsymbol{V}}_{ym}\boldsymbol{x}_{yi} + \tfrac{1}{2}\hat{\gamma}_{ym}\hat{\sigma}^2_{ym}\right) \quad 4.1$$

where we assume that $\hat{\boldsymbol{V}}_{ym}, \hat{\gamma}_{ym}$ and $\hat{\sigma}^2_{ym}$ are sufficiently good estimators for $\boldsymbol{V}_{ym}$, γ_{ym} and σ^2_{ym} that no further bias correction is required (see the discussion after equation 3.6).

The index movement within year y is $\bar{P}_{Dym}/\bar{P}_{Dy01}$ and the index relative to February 2002, is:

$$I_{Dym} = 100\frac{\bar{P}_{D0213}}{\bar{P}_{D0202}}\frac{\bar{P}_{D0313}}{\bar{P}_{D0301}}\cdots\frac{\bar{P}_{D,y-1,13}}{\bar{P}_{D,y-1,1}}\frac{\bar{P}_{Dym}}{\bar{P}_{Dy01}} = 100\left(\frac{\bar{P}_{D0213}}{\bar{P}_{D0202}}\prod_{z=03}^{y-1}\frac{\bar{P}_{Dz13}}{\bar{P}_{Dz01}}\right)\frac{\bar{P}_{Dym}}{\bar{P}_{Dy01}} \quad 4.2$$

where month 13 of year y-1 = month 1 of year y .

The index may alternatively be expressed as:

$$I_{Dym} = 100\frac{1}{\bar{P}_{D0202}}\frac{\bar{P}_{D0213}}{\bar{P}_{D0301}}\cdots\frac{\bar{P}_{D,y-2,13}}{\bar{P}_{D,y-1,01}}\frac{\bar{P}_{D,y-1,13}}{\bar{P}_{Dy01}}\bar{P}_{Dym} = 100\left(\prod_{z=02}^{y-1}\frac{\bar{P}_{Dz13}}{\bar{P}_{D,z+1,01}}\right)\frac{\bar{P}_{Dym}}{\bar{P}_{D0202}} \quad 4.3$$

The ratios $\bar{P}_{Dz13}/\bar{P}_{D,z+1,1}$ are link factors, which ensure continuity of the chain-linked index by correcting for the change, at every January link month, in the weights and model used in the index.

Note that the constraint $\sum_{i\in D} w_{Dyi} = 1$ is only necessary for the calculation of a weighted average price. Equation 4.2 expresses the house price index as a product of ratios of weighted averages. Each ratio contains the same weights in both numerator and denominator. So, for the house price index, only the relative sizes of the weights are important and the constraint $\sum_{i\in D} w_{Dyi} = 1$ is not necessary.

5. Weighted Regressions

Sometimes, the data we receive from lenders are incomplete, with missing values for some covariates. Values may be missing for all values of a covariate (or covariates) from a single lender or only for a few observations. Rather than reject observations with missing values, we have chosen to keep them in the regression analysis with a special value for each covariate to indicate a missing value.

In doing so, we are assuming that, for a given set of known covariates, the missing values are "missing at random". That is, the true values for the missing covariates may be represented as a random realisation from the population distribution of these covariates, for the given set of known covariates. This is equivalent to the more general assumption that covariates not included in the model at all (for example: garage; state of repair of the property) are "missing at random" and may be represented by a random error term. Provided this assumption is valid, there is no bias in the parameters estimated in this way.

However, the assumption of constant variance for the error terms is no longer valid because, clearly, observations with missing values will have greater variance than observations with complete covariate data.

To resolve this problem, we assume that the error variance for observations with missing data is equal to the basic error variance σ^2_{ym} for observations with complete data multiplied by a factor greater than 1. The non-uniform variances can then be accommodated by applying a weighted regression, using observation weights equal to the reciprocals of estimates of these factors.

For any given combination of missing covariates, we can estimate these weights, which are less than or equal to 1, from regressions based on complete data only. The estimated observation weight, for a given set of missing covariate values, is calculated as the ratio of the mean squared error from the model including all covariates to the mean squared error from the model including only the non-missing covariates.

It may be desirable to re-calculate these weights every month, using the complete data available for that month. However, this would be time-consuming, when the main operational requirement is to validate the data and produce the current month's index values. We believe that these weights will be stable over time and that small errors in these weights would not have any noticeable impact on the estimated indices. However, these observation weights will be re-assessed annually, to ensure that long-term changes to the factors affecting house prices are picked up and to accommodate any changes to the covariates included in the model.

Using weighted regressions, means that the analyses presented in sections 1 to 3 above need to be modified appropriately. There are no modifications to the underlying principles, because weighted regressions can easily be accommodated by appropriate transformations to the dependent variable and covariates. The required modifications and transformations are covered in standard text books (Johnson, 1972, for example).

In summary, the following equations from sections 1 to 3 need to be modified as shown, to accommodate the application of weighted regressions.

Sections 1 and 2

The original model in section 1 may be re-written as:

$$L_{ymik} = \ln\left(P_{ymik}\right) = \boldsymbol{x}'_{yi}\boldsymbol{\beta}_{ym} + \varepsilon^{*}_{ymik} \qquad 5.1 \equiv 2.1a$$

where the modified error terms have the properties:

$$E\left[\varepsilon^{*}_{ymik}\right] = 0; \quad Var\left[\varepsilon^{*}_{ymik}\right] = \frac{\sigma^{2}_{ym}}{\lambda_{ymik}}; \quad Cov\left[\varepsilon^{*}_{ymik}, \varepsilon^{*}_{znjl}\right] = 0 \quad \text{for } ymik \neq znjl \qquad 5.2 \equiv 2.2a$$

with λ_{ymik} being the appropriate weight for observation *ymik*.

It is possible to transform this model into a model whose error terms have constant variance by transforming the dependent variable and covariates, as follows:

$$\sqrt{\lambda_{ymik}}\,L_{ymik} = \sqrt{\lambda_{ymik}}\,\ln\left(P_{ymik}\right) = \sqrt{\lambda_{ymik}}\,\boldsymbol{x}'_{yi}\boldsymbol{\beta}_{ym} + \varepsilon_{ymik} \qquad 5.3 \equiv 2.1b$$

so that the error term $\varepsilon_{ymik} = \sqrt{\lambda_{ymik}}\,\varepsilon^{*}_{ymik}$ has the homoscedastic property of equations 2.2 .

Section 3

$$E\left[\boldsymbol{\varepsilon}^{*}_{ym}\right] = \boldsymbol{0}; \quad Var\left[\boldsymbol{\varepsilon}^{*}_{ym}\right] = \sigma^{2}_{ym}\boldsymbol{\Omega}_{ym} = \sigma^{2}_{ym}\boldsymbol{W}^{-1}_{ym} \qquad 5.4 \equiv 3.1a$$

where $\boldsymbol{W}_{ym}$ is a diagonal matrix of the weights $\{\lambda_{ymik}\}$.

Using weighted least squares estimation, the estimator for $\boldsymbol{\beta}_{ym}$ is:

$$\hat{\boldsymbol{\beta}}_{ym} = \left(\boldsymbol{X}'_{ym}\boldsymbol{W}_{ym}\boldsymbol{X}_{ym}\right)^{-1}\boldsymbol{X}'_{ym}\boldsymbol{W}_{ym}\boldsymbol{l}_{ym} \qquad 5.5 \equiv 3.2a$$

with: $Var\left(\hat{\boldsymbol{\beta}}_{ym}\right) = \sigma^{2}\left(\boldsymbol{X}'_{ym}\boldsymbol{W}_{ym}\boldsymbol{X}_{ym}\right)^{-1} = \boldsymbol{V}_{ym}$ (for notational convenience) $\qquad 5.6 \equiv 3.3a$

(see, for example, Johnston, 1972, chapter 7).

References

Cook, J, and Weisberg, S. (1982) *Residuals and Influence in Regression*, Chapman & Hall, London

Johnston, J. (1972) *Econometric Methods*, 2nd ed., McGraw-Hill, Tokyo

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* , **6**, 461-464.

Shi, P., Tsai, C-L. (2002), Regression model selection - a residual likelihood approach, *J. R. Statist. Soc.* B, **64**, *Part 2*, 237-252

Stuart, A. and Ord, J.K. (1987) *Kendall's Advanced Theory of Statistics*, 5th ed., Vol.1, Griffin, London

Appendix

Estimation of Cumulants

1. Introduction

The model described in section 1 assumes that the error terms are independent and identically distributed but not that they are Normally distributed. As discussed in section 2, the higher order moments, in the form of cumulants, are relevant for the unbiased estimation of average house prices. This appendix develops estimators for the third and fourth cumulants.

For clarity of exposition, the suffices relating to year, month and cell will be omitted. Suffices will define only the observation number, out of n observations using a model fitted to p parameters. For the House Price Index, $n \approx 20{,}000$ and $p \approx 200$.

The weighted residual from the model for observation i is defined as:

$$e_i = \sqrt{\lambda_i}\hat{r}_i = \sqrt{\lambda_i}\left(l_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right) = \sqrt{\lambda_i}l_i - \sqrt{\lambda_i}\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$$

Standard regression theory states that the vector of residuals $\boldsymbol{e}$ may be expressed as a linear function of the vector of independent, identically distributed error terms $\boldsymbol{\varepsilon}$ (see, for example, Johnston, 1972): $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon}$, where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{H}$ is the so-called "hat" matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

In particular:

$$e_i = \sum_{j=1}^{n} M_{ij}\varepsilon_j = \sum_{j=1}^{n}\left(I_{ij} - H_{ij}\right)\varepsilon_j$$

Note that $I_{ij} = 1$, $I_{ij} = 0$ for $i \neq j$ and $(I_{ij})^r = I_{ij}$ for all i, j, r .

Also:

$$H_{ii} \leq 1, \quad tr\left(\boldsymbol{H}\right) = \sum_{i=1}^{n} H_{ii} = p \quad \text{and} \quad \sum_{j=1}^{n} H_{ij}^2 = H_{ii}$$

because $\boldsymbol{H}$ is a symmetric, idempotent matrix of rank p (see, for example, Cook & Weisberg, 1982).

2. Estimating the Error Variance σ^2

We illustrate our approach to developing estimators for cumulants by applying it to produce the standard unbiased estimator for the error variance σ^2. We have:

$$E\left[\sum_{i=1}^{n} e_i^2\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}\varepsilon_j \sum_{k=1}^{n} M_{ik}\varepsilon_k\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} M_{ij}M_{ik}E\left[\varepsilon_j\varepsilon_k\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2\sigma^2$$

$$= \sigma^2\sum_{i=1}^{n}\sum_{j=1}^{n}\left(I_{ij} - H_{ij}\right)^2$$

$$= \sigma^2\sum_{i=1}^{n}\sum_{j=1}^{n}\left(I_{ij} - 2I_{ij}H_{ij} + H_{ij}^2\right)$$

$$= \sigma^2\sum_{i=1}^{n}\left(1 - 2H_{ii} + H_{ii}\right)$$

$$= \sigma^2\left(n - p\right)$$

Hence we have the standard, unbiased estimator for the error variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{\left(n - p\right)}$$

3. Estimating the Third Cumulant κ_3

$$E\left[\sum_{i=1}^{n} e_i^3\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}\varepsilon_j \sum_{k=1}^{n} M_{ik}\varepsilon_k \sum_{l=1}^{n} M_{il}\varepsilon_l\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} M_{ij}M_{ik}M_{il}E\left[\varepsilon_j\varepsilon_k\varepsilon_l\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^3\mu_3$$

because the independence of the ε_i means that $E[\varepsilon_j\varepsilon_k\varepsilon_l]=0$ if any subscript j, k or l is different from the other two.

Hence, noting also that $\mu_3=\kappa_3$:

$$E\left[\sum_{i=1}^{n} e_i^3\right] = \kappa_3 \sum_{i=1}^{n}\sum_{j=1}^{n}\left(I_{ij} - H_{ij}\right)^3$$

$$= \kappa_3 \sum_{i=1}^{n}\sum_{j=1}^{n}\left(I_{ij} - 3I_{ij}H_{ij} + 3I_{ij}H_{ij}^2 - H_{ij}^3\right)$$

$$= \kappa_3 \left\{\sum_{i=1}^{n}\left(1 - 3H_{ii} + 3H_{ii}^2\right) - \sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}^3\right\}$$

$$= \kappa_3 \left\{(n-3p) + 3\sum_{i=1}^{n} H_{ii}^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}^3\right\}$$

The mean value of H_{ii} is p/n. If the absolute deviations of the H_{ii} from the mean p/n are mainly small (that is, there are not too many large values of H_{ii}), then each H_{ii} will be of the order p/n and $\frac{p^2}{n} \lessapprox \sum_{i=1}^{n} H_{ii}^2 \le p$. In practice, for the house price data, $\sum_{i=1}^{n} H_{ii}^2 \approx 4$. This is very small relative to $(n\text{-}3p)$ and may be neglected.

Using the same reasoning, the mean value of H_{ij}^2 $(i\neq j)$ is approximately $\frac{p - p^2/n}{n^2 - n} \approx \frac{p}{n^2}$. Again, if the deviations of $|H_{ij}|$ from the mean absolute value $(\sqrt{p})/n$ are sufficiently small, we would expect $\sum_{i=1}^{n}\sum_{j=1}^{n}\left|H_{ij}^3\right| \approx \frac{p^{1.5}}{n}$. In fact, some of the H_{ij} may be negative and $\sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}^3$ will be smaller than this. So this term, too, may be neglected.

We therefore have the following, approximately unbiased estimator for the third cumulant:

$$\hat{\kappa}_3 = \frac{\sum_{i=1}^{n} e_i^3}{n - 3p}$$

4. Estimating the Fourth Cumulant κ_4

$$E\left[\sum_{i=1}^{n} e_i^4\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}\varepsilon_j \sum_{k=1}^{n} M_{ik}\varepsilon_k \sum_{l=1}^{n} M_{il}\varepsilon_l \sum_{m=1}^{n} M_{im}\varepsilon_m\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} M_{ij}M_{ik}M_{il}M_{im}E\left[\varepsilon_j\varepsilon_k\varepsilon_l\varepsilon_m\right]$$

The independence of the ε_i means that $E[\varepsilon_j\varepsilon_k\varepsilon_l\varepsilon_m]=0$ if any subscript j, k, l or m is different from all the other three. So $E[\varepsilon_j\varepsilon_k\varepsilon_l\varepsilon_m]$ is not zero only if pairs of subscripts are equal, that is:

$j=k$ and $l=m$ ($j\neq l$); $j=l$ and $k=m$ ($j\neq k$); $j=m$ and $k=l$ ($j\neq k$); or $j=k=l=m$.

The first three conditions are equivalent, corresponding merely to rotation of subscripts, so:

$$E\left[\sum_{i=1}^{n} e_i^4\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^4 E\left[\varepsilon_j^4\right] + 3\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} M_{ij}^2 M_{ik}^2 E\left[\varepsilon_j^2\varepsilon_k^2\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^4\mu_4 + 3\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} M_{ij}^2 M_{ik}^2 E\left[\varepsilon_j^2\right]E\left[\varepsilon_k^2\right] \quad \text{(from the independence of } \varepsilon_j \text{ and } \varepsilon_k \text{)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^4\left(\kappa_4 + 3\sigma^4\right) + 3\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} M_{ij}^2 M_{ik}^2\sigma^4$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^4\kappa_4 + 3\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} M_{ij}^2 M_{ik}^2\sigma^4$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^4\kappa_4 + 3\sum_{i=1}^{n}\left(\sum_{j=1}^{n} M_{ij}^2\right)^2\sigma^4$$

Hence:

$$E\left[\sum_{i=1}^{n} e_i^4\right] = \kappa_4 \sum_{i=1}^{n}\sum_{j=1}^{n}\left(I_{ij} - H_{ij}\right)^4 + 3\sigma^4 \sum_{i=1}^{n}\left\{\sum_{j=1}^{n}\left(I_{ij} - H_{ij}\right)^2\right\}^2$$

$$= \kappa_4 \sum_{i=1}^{n}\sum_{j=1}^{n}\left(I_{ij} - 4I_{ij}H_{ij} + 6I_{ij}H_{ij}^2 - 4I_{ij}H_{ij}^3 + H_{ij}^4\right) + 3\sigma^4 \sum_{i=1}^{n}\left\{\sum_{j=1}^{n}\left(I_{ij} - 2I_{ij}H_{ij} + H_{ij}^2\right)\right\}^2$$

$$= \kappa_4 \left\{\sum_{i=1}^{n}\left(1 - 4H_{ii} + 6H_{ii}^2 - 4H_{ii}^3\right) + \sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}^4\right\} + 3\sigma^4 \sum_{i=1}^{n}\left\{1 - H_{ii}\right\}^2$$

$$= \kappa_4 \left\{n - 4p + 6\sum_{i=1}^{n} H_{ii}^2 - 4\sum_{i=1}^{n} H_{ii}^3 + \sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}^4\right\} + 3\sigma^4 \sum_{i=1}^{n}\left(1 - 2H_{ii} + H_{ii}^2\right)$$

$$= \kappa_4 \left\{n - 4p + 6\sum_{i=1}^{n} H_{ii}^2 - 4\sum_{i=1}^{n} H_{ii}^3 + \sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}^4\right\} + 3\sigma^4 \left(n - 2p + \sum_{i=1}^{n} H_{ii}^2\right)$$

Using the same arguments presented above for the third cumulant κ_3, we may ignore the summation terms in the line above to obtain the approximation:

$$E\left[\sum_{i=1}^{n} e_i^4\right] \approx \kappa_4 \left\{n - 4p\right\} + 3\sigma^4 \left(n - 2p\right)$$

from which we have the following, approximately unbiased estimator for the fourth cumulant κ_4:

$$\hat{\kappa}_4 = \frac{\sum_{i=1}^{n} e_i^4 - 3\hat{\sigma}^4\left(n - 2p\right)}{\left(n - 4p\right)}$$

Note that $\hat{\sigma}^4 = \left(\hat{\sigma}^2\right)^2$ is a biased estimator of σ^4 but the bias is of the order κ_4/n, which is negligible to the degree of approximation used here.