BLS WORKING PAPERS

U.S. Department of Labor
U.S. Bureau of Labor Statistics
Office of Prices and Living Conditions

Improving the CPI's Age-Bias Adjustment: Leverage, Disaggregation and Model Averaging

Joshua Gallin, Board of Governors of the Federal Reserve System
Randal Verbrugge, U.S. Bureau of Labor Statistics

Working Paper 411
October 2007

Improving the CPI's Age-Bias Adjustment: Leverage, Disaggregation and Model-Averaging

Joshua Gallin
Board of Governors of the Federal Reserve System
Washington, DC
Email: Joshua.H.Gallin@frb.gov

Randal J. Verbrugge
Division of Price and Index Number Research
Bureau of Labor Statistics
2 Massachusetts Ave. NE
Washington, DC 20212
Email: verbrugge.randal@bls.gov

First Version: December 6, 2004

This version completed on: October 17, 2007

JEL Classification Codes: E31; C81; C82; R31; R21; O47

Keywords: depreciation; hedonics; model averaging; inflation; CPI bias.

Acknowledgement 1 *We thank David Johnson, Louise Campbell, Tim Erickson, John Greenlees, Johannes Hoffmann, Greg Kurtzon, Frank Ptacek and Elliot Williams, and participants at the DPINR research seminar and the 2006 Western Economics Association Conference. All errors, misinterpretations and omissions are ours. All the analysis, views, and conclusions expressed in this paper are those of the authors; they do not reflect the views or policies of the Bureau of Labor Statistics or the views of other BLS staff members; neither do they indicate concurrence by other members of the research staff of the Board of Governors, or by the Board of Governors itself.*

Abstract

As a rental unit ages, its quality typically falls; a failure to correct for this would result in downward bias in the CPI. We investigate the BLS age bias imputation and explore two potential categories of error: approximations related to the construction of the age bias factor, and model misspecification. We find that, as long as one stays within the context of the current official regression specification, the approximation errors are innocuous. On the other hand, we find that the official regression specification – which is more or less of the form commonly used in the hedonic rent literature – is severely deficient in its ability to match the conditional log-rent vs. age relationship in the data, and performs poorly in out-of-sample tests. It is straightforward to improve the specification in order to address these deficiencies.

However, basing estimates upon a single regression model is risky. Age-bias adjustment inherently suffers from a general problem facing some types of hedonic-based adjustments, which is related to model uncertainty. In particular, age-bias adjustment relies upon specific coefficient estimates, but there is no guarantee that the true marginal influence of a regressor is being estimated in any given model, since one cannot guarantee that the Gauss-Markov conditions hold. To address this problem, we advocate the use of model averaging, which is a method that minimizes downside risks related to model misspecification and generates more reliable coefficient estimates. Thus, after selecting several appropriate models, we estimate age-bias factors by taking a trimmed average over the factors derived from each model. We argue that similar methods may be readily implemented by statistical agencies (even very small ones) with little additional effort.

We find that, in 2004 data, BLS age-bias factors were too small, on average, by nearly 40%. Since the age bias term itself is rather small, the implied downward-bias of the aggregate indexes is modest. On the other hand, errors in particular metropolitan areas were much larger, with annual downward-bias as large as 0.6%.

1 Introduction

Housing costs are a substantial part of most American's monthly outlays. As a result, these costs account for about one-third of the total weight of the Consumer Price Index (CPI). There are two major components of these shelter costs. First, there is tenant's rent, covering the shelter expenditures of renters. Second, there is owner's equivalent rent (OER), covering the shelter expenditures of owners. (The rental equivalence method – which abstracts from the highly-volatile, difficult-to-measure, financial-asset aspect of homeownership – is probably the best of the available methods for estimating changes in homeowner shelter costs. For details, see Ptacek and Baskin 1996, Diewert 2003, Poole, Ptacek and Verbrugge 2005, and Verbrugge 2007.)

Since shelter expenditures have such an enormous weight in the CPI, accurate measurement of shelter costs is crucial to obtaining an accurate measurement of the overall inflation experienced by the average US consumer. The measurement goal for the shelter components of the CPI is cost changes in *constant-quality* housing units. To approximate quality-adjusted price changes, the Bureau of Labor Statistics (BLS) collects rental price data from a sample of housing units over time, and makes adjustments for changes in observed physical characteristics (such as the number of rooms), and for aging.[1] Failing to adjust for aging would introduce a downward bias into the CPI, since housing units deteriorate over time; an unchanged rent on unit which has aged represents a price increase, since the same amount of money is purchasing a lower-quality good. (Of course, renovation and remodeling can temporarily reverse the deterioration experienced by a particular unit; the adjustment for aging is based upon the *net* effect of aging and renovation within a metropolitan area.)

The BLS adjustments for aging are based on Randolph (1988a), and involve scaling observed rents by "age-bias" factors that are based on a hedonic model for housing rents.[2] For any given year, this adjustment is fairly modest; still, its impact over many years is nontrivial. In this paper, we describe several shortcomings in the way the BLS specifies its hedonic model, and show that these shortcomings can have significant effects on reported CPI housing indexes.

The BLS estimates its hedonic model every year using cross-section data from its housing unit

[1] See the *BLS Handbook of Methods*.
[2] see also Lane, Randolf, and Berenson (1988)

sample merged with data from the decennial Census. The model assumes that the log rent for a given housing unit depends on the unit's age, age-squared, age interacted with several housing-unit characteristics, and a large set of controls for other physical and neighborhood characteristics. We use the same data to illustrate two types of shortcomings of the BLS methodology. First, as described below, the BLS makes two approximations in constructing aging factors, which turn out to impart a bias to these estimates. Second, the BLS model specification suffers from two problems: it allows a small number of extremely old units to have a large effect on the estimated age-bias factors, and it is too restrictive in that it does not take full advantage of the available geographic information. In particular, the BLS hedonic model has only five age-related terms, and estimates its hedonic model separately only for each of the four Census regions, thereby ignoring the diversity across statewide or metropolitan housing markets. (A third possible shortcoming is a potential bias related to a confounding of the effects of historical depreciation of *surviving* units with average annual depreciation of *all* units, and the inability of the BLS procedure to control for unit-specific characteristics. We investigate this set of topics in a companion paper, Gallin and Verbrugge 2007.)

Age-bias adjustment also inherently suffers from a general problem facing some types of hedonic-based adjustments: it relies upon specific coefficient estimates, but there is no guarantee that the true marginal influence of a regressor is being estimated in any particular model, since one cannot guarantee that the Gauss-Markov conditions hold. The implied estimated marginal influence of a given variable can differ dramatically across models which otherwise appear roughly equivalent in terms of their complexity, their fit to the data, their out-of-sample predictive ability, and so on. (Coefficient estimates often change when the model changes, which is why empirical work often includes tables of regression results for different models.) We argue that the potential for such variability is a compelling argument for a model-averaging approach, which reduces the risk of choosing a model whose coefficient estimates are far from the true marginal effects. Simple variants of model averaging can be implemented with very little additional work.

We use the BLS and Census data to show that simple remedies for these shortcomings alter estimates of age-bias factors. Our estimated age-bias factors are on average almost 40% larger than, and often *quite* different from, those produced by the BLS methods. The effects turn out to be

relatively small at the national level. In contrast, the resulting estimates of the growth rate of the CPI for housing can differ importantly from the baseline BLS approach in many metropolitan areas. In particular, these estimates can be altered by more than 0.6%,[3] so that – for example – estimated rent inflation in a metropolitan area might increase from 1.5% to 2.1%, easily large enough to alter the overall inflation rate in the metropolitan area. This in turn can be of major local significance; for example, Colorado's Amendment 42, which passed in 2006, indexes Colorado's minimum wage to Denver's CPI.

2 Age-Bias Adjustment of the CPI for Housing

2.1 Description

The CPIs for renter- and owner-occupied housing are meant to measure "price" changes (in this case, rent changes) for the service flow from a constant-quality unit of housing. The BLS uses three methods to control for changes in quality. First, estimates of rent growth are based on a panel sample, so the same units are tracked over time.[4] Second, the BLS makes adjustments to account for major changes in a housing unit's physical characteristics, such as the number of rooms. Third, the BLS corrects for so-called "age-bias" by scaling observed rents by an age-bias factor that controls for changes in the quality of a housing unit that owes to aging. The focus of this paper is to investigate the BLS procedure for adjusting for aging bias. In doing so, we illustrate some potential shortcomings of typical procedures for specifying and using hedonic models.

The BLS constructs its index for Rent (or OER) for an area k, I^k, using a "rent relative" approach. In particular,

$$I_t^k = I_{t-1}^k * R_t^k$$

where R is the rent relative, and t indexes months. As explained in Section 3 below, the BLS reprices the housing units in their sample only every six months. Accordingly, the rent relative –

[3] The metropolitan-area aging bias estimates depend to an appreciable extent upon modeling choices of the sort we investigate. The estimates we offer here are conservative, and might understate true changes to the baseline BLS approach.

[4] Sample attrition over time is unavoidable, as units are demolished, or as units become vacant and then tenants are replaced by unresponsive tenants. BLS procedures account for such unobserved rent changes; see Ptacek and Baskin (1996) and Crone, Nakamura and Voith (2006).

which is used to move the index in the current month t – is defined as

$$R_t^k = \left(\frac{\sum w_i \text{rent}_{i,t}}{\sum w_i \text{rent}_{i,t-6} e^{F_{i,t}^k}} \right)^{\frac{1}{6}} \tag{1}$$

where w is an expenditure weight and $F_{i,t}^k$ is the age-bias factor; i indexes housing units.[5] The age-bias factor, roughly speaking, adds six months of aging to the $t-6$ unit, in order to compute inflation based upon constant-quality units. The expenditure weights for the Rent relative differ from those for the OER relative, since – for example – the OER expenditure weights are zero on all rent-control units. (For further discussion on the OER approach to pricing shelter service inflation for homeowners, see Poole, Ptacek and Verbrugge, 2005.)

The age-bias factor F^k is based on a hedonic regression for rent. The BLS model is of the general form

$$\ln rent_{i,t} = \alpha_t + \gamma_1 age_{i,t} + \gamma_2 age_{i,t}^2 + \widetilde{\gamma}_3 \widetilde{z}_{i,t} age_{i,t} + \widetilde{\beta} \widetilde{X}_{i,t} + u_{i,t} \tag{2}$$

where $\widetilde{X}_{i,t}$ and $\widetilde{z}_{i,t}$ are each vectors. In particular, $\widetilde{X}$ is a vector that includes over 20 measures of unit-level characteristics (such as number of rooms, and whether the structure is detached or multi-unit), dummy variables indicating the size of the metropolitan area (termed a "Primary Sampling Unit," or PSU), and Census neighborhood variables (such as percent of population that is under the poverty line). The vector $\widetilde{z}_{i,t}$ consists of three variables: the number of rooms, a dummy variable indicating if a unit is a detached unit, and a dummy variable indicating if a unit is aged 85 years or more. Thus there are five terms related to age.[6] While the BLS expends considerable effort in determining the correct age of each unit, in some cases the age of the unit must be estimated, based upon (for example) knowledge of the decade in which the structure was built. If there is no reliable information on the age of a unit at all, such units are excluded from the regression. The BLS estimates the regression coefficients separately for each Census region (*BLS*, 2006). The data used are from July-December of year t.

The CPI's age-bias factor for area k in the following year is equal to the partial derivative of equation (2) with respect to age, evaluated at the area-level averages for age, $\overline{\text{age}_k}$ and the interaction terms, $\overline{\widetilde{z}_k}$. In other words, the common 6-month age-bias factor for all housing units in

[5] We ignore many technical details, such as nonresponse adjustment and utilities adjustment for OER, the latter of which is studied in Verbrugge (2007). For more details, see Ptacek and Baskin (1996).

[6] These terms are called "depreciation terms" in Lane, Randolph, and Berenson (1988)

area k is given by

$$F_k^{BLS} = \frac{1}{2}(\widehat{\gamma}_1 + 2\widehat{\gamma}_2\overline{\text{age}_{k,t}} + \widehat{\widetilde{\gamma}}_3\overline{\widetilde{z}_{k,t}}) \tag{3}$$

where division by 2 converts the annual aging factor into a semi-annual factor.

2.2 Discussion

Adjusting OER The age-bias factor is derived from a hedonic regression on rental units, but is applied in the computation of both the Rent and OER indexes. It is sometimes argued that, since owner units likely depreciate at a different rate than rental units, it is erroneous to use an age-bias factor which has been estimated from rental-market data. However, this objection is invalid. The measurement goal for OER is inflation in the shelter service price. Since OER is constructed from inflation in market rents – which by definition, are from rented units – the aging correction required to properly remove the effects of depreciation on rented units must also be estimated from rented-unit data. What remains after the correction is constant-quality shelter-service inflation, which is precisely the measurement goal.

Approximation errors Computing age-bias factors using (3) involves making two approximations, which we discuss and investigate in Section 5 below. These approximations are not perfect and impart biases into (2); but it turns out that these are of a small magnitude and largely offsetting, as long as one remains in the context of the BLS model (2).

Coefficient-estimates versus true marginal effects The intention of age-bias adjustment is to adjust for the marginal effect of age on rent; what is ultimately required is an accurate estimate of this true marginal effect. In practice, the age-bias factor (3) is constructed using particular coefficient estimates from a regression model. Hence, age-bias adjustment accuracy requires that particular coefficient estimates accurately estimate the true marginal effect of age.[7] But in any given model, one cannot guarantee that the Gauss-Markov conditions hold; so the coefficient estimates $\widehat{\gamma}_1, \widehat{\gamma}_2$, and $\widehat{\widetilde{\gamma}}_3$ might imply marginal effects of age that are quite different from reality (despite the

[7] Unfortunately, unlike many other applications of hedonics, age-bias adjustment cannot make use of the extremely useful fact that a regression model like (2) can deliver unbiased predictions for missing left-hand-side variables – given a full set of right-hand-side variables whose joint distribution is the same as the variables used to estimate the model. See Erickson and Pakes (2007) for an unbiased-prediction application to quality change in television data.

fact that, *in the context of the regression model being estimated*, the coefficient estimates could well be accurately capturing the marginal relationship of age to the (conditional) expectation of the dependent variable).

A hypothetical case, namely quality-adjusting for newly-installed air-conditioning in a Los Angeles apartment, illustrates this point. Adding air-conditioning to a non-air-conditioned apartment is clearly a quality increase to that apartment. If this unit were in the BLS housing sample, then a quality adjustment (or a "structural change adjustment" in the parlance of the BLS) would be required in order to avoid bias in the shelter indexes. A hedonics-based adjustment would rely upon the estimated coefficient on the relevant air-conditioning dummy variable. (In the BLS specification, there are three air-conditioning dummy variables: "central," "window," and "other.") But suppose that the window air-conditioning coefficient estimate was negative and statistically significant at conventional levels. This would undoubtedly reflect overall correlations in the data – for example, window air-conditioning being negatively correlated with an unobserved quality variable. But the implied quality adjustment is then negative – even though it is clear that this is a quality *improvement*, with an *upward* impact on rent.[8]

Specification error can lead to biased coefficient estimates; careful attention to specification is simply good practice. But specification testing does not solve the more general problem. (And unfortunately, applying conventional specification-search procedures can readily yield models which imply *less* reliable age-bias factors.) Since any particular regression model may yield unreliable coefficient estimates, this strongly supports the practice of model averaging: selecting several repectable regression models, and averaging the age-adjustments estimated by each. Below, we discuss the rationale for model averaging in more detail, and suggest an appropriate and simple averaging method.

Sign of the age-bias adjustment All structures deteriorate over time. Most receive maintenance that helps offset the deterioration, and some receive major improvements that temporarily reverse the deterioration. Furthermore, certain age-groups or vintages of units might become *more* desirable over time, which could offset or reverse the otherwise downward effect of aging on rent.

[8] This is not a blanket criticism of the use of hedonics. Indeed, hedonics are an elegant and rigorous solution to many challenging problems in price indexes. In other cases – such as adjusting for the effect of aging – it is the only game in town, and far better than doing nothing.

However, all housing is eventually torn down or completely renovated, which implies that the *average* housing unit depreciates (Lane, Randolph, and Berenson, 1988). Thus, we expect that PSU-average age-bias factors will be negative.

3 The data

The data used are neighborhood data from the decennial 2000 Census, and confidential BLS rental housing microdata from July-December 2004. As the Census data is well-known and described elsewhere, we here describe the BLS data.

Decisions regarding the BLS methodology for rental housing sampling are described in Ptacek and Baskin (1996). In brief, for each of the metropolitan areas (Primary Sampling Units, or PSU's) in the BLS sample, the BLS randomly selects a geographically-diverse set of rental housing units, via a geographic stratification procedure. In the initial data collection steps, the BLS collects a large amount of information about each unit, such as its age, structural characteristics (e.g., "located within a multi-unit building with an elevator," "detached unit," etc.), number of bedrooms and bathrooms, utilities (including whether utilities are included in the rent), and so on. The housing sample is divided into six panels; that is, each unit is placed into one of six panels. Rent price data on all the units in a particular panel are collected in the same month, and then – given that a typical unit experiences a rent price change every twelve months (see Crone, Nakamura and Voith (2006) – not again until six months later. Each panel is thus priced twice a year; for example, panel 1 is priced in January and July, panel 2 in February and August, and so on.

A typical unit remains in the sample for many years. The BLS data we use are from the second half of 2004.[9] Table 1 lists the BLS microdata variables used, along with some descriptive statistics.

[9] 2003 data yield results which, if anything, are more striking.

Rent sample distribution statistics

	1%	25%	50%	75%	99%
log rent	5.4	6.2	6.5	6.8	7.7
age	4	23	35	54	128
bedrooms	0	1	2	2	4
bathrooms	1	1	1	2	3
other rooms	1	2	2	3	4

% of rental sample featuring:

single family detached	21%	electric heat	41%	central A/C	44%
duplex/townhouse	18%	gas heat	50%	window A/C	15%
multi-unit w/ elevator	9%	other heating fuel	1%	other A/C	11%
multi-unit w/out elevator	50%	heat included	18%		
mobile home	2%	electricity included	7%		

Table 1

4 Empirical Strategy

We investigate two potential types of shortcomings in BLS methodology. The first relates to the aforementioned approximation errors in estimating age-bias factors. Below, we demonstrate the shortcomings of these assumptions: they impart a bias on the estimated aging factors. (This bias turns out to be fairly small for the BLS hedonic model, though we demonstrate that it becomes a lot more problematic under other empirical specifications.)

The other set of potential shortcomings relates to model specification issues. The first of these relates to overly-influential observations. The distribution of age across units is strongly skewed to the right. This suggests that extremely old units could well have high leverage, i.e. that they have a inappropriately-large impact on the coefficients related to age, since the quadratic specification estimated by OLS will heavily penalize large errors on old units. We investigate this issue, examining in a simple way both the leverage of old units as a group, as well as the extent to which coefficient

estimates are altered by inclusion or exclusion of very old units. (A dummy-variable approach is already in use; we demonstrate that it does not solve the problem. We also investigate the use of a specification which is based upon $\ln(age)$.)

The second specification issue relates to whether or not the BLS model is unduly restrictive. We investigate two types of restrictions. First, the BLS model restricts coefficient estimates to be identical within Census regions; is further disaggregation warranted? Second, the BLS model has only five age-related terms, assumes that age effects are quadratic (aside from linear interactions), and implicitly imposes a common average log-rent across PSU's. Are other specifications superior?

As we argue that basing one's estimates upon a single model is risky, we advocate a simple form of model-averaging.

We use a straightforward metric for determining whether any particular shortcoming is problematic: to what extent is the rent or OER inflation rate impacted by a particular potential remedy? This is, after all, the bottom line. However, applying this metric is not entirely straightforward, since deficiencies can interact; for example, using a piecewise-linear or higher-order polynomial in age implies that a PSU-average approach can impart an significant bias upon estimated age-bias factors. Furthermore, although we attempt to replicate BLS methods in estimating age-bias factors, we do not have access to BLS's full set of production programs, and therefore cannot exactly replicate the BLS estimates. We perform our own estimation of both these factors and our alternative factors. In this way, any imperfection in our procedure for estimating aging factors will likely net out.[10]

In particular, we compute our metric as follows. We estimate a baseline model in which we mimic the BLS hedonic model for 2004, which yields baseline age-bias factors, $\widehat{F}^{BLS}_{i,2004}$. We compare alternative age-bias factors, denoted $F^{ALT}_{i,t}$, to this baseline. Using (1), it is straightforward to deduce that the revision to this relative is given by

$$\frac{R^{ALT}_t}{R^{BLS}_t} = \frac{\sum w_i r_{i,t-6} e^{\widehat{F}^{BLS}_{i,t}}}{\sum w_i r_{i,t-6} e^{F^{ALT}_{i,t}}}. \tag{4}$$

For example, if (4) equals 1.02, this implies that our alternative factors would have generated an

[10] Our estimates of the BLS factors are very similar to the actual BLS factors; the correlation coefficient across the 87 PSUs is about .95. Thus, we are confident that our strategy yields reliable results.

inflation estimate that was 2% larger than the official estimate. As noted above, we approximate $F_{i,t}^{BLS}$ with our estimate, $\widehat{F}_{i,t}^{BLS}$. Notice that the size of the revision depends upon the weights w_i, which differ across units and across indexes (Rent or OER).

5 Approximation Errors

As noted above, computing age-bias factors as in (3) involves making two approximations. First, because the age-bias factor is based on the partial derivative of equation (2), it can only precisely represent the effect of an *infinitesimal* change in age, rather than the effect due to a discrete change in age. Second, (3) generates a common age-bias factor across all units within the the same PSU, via the use of PSU-averages in the formula.

It is straightforward to show that the correct six-month unit-i-specific age-bias term implied by model (2) is

$$F_i = \frac{\widehat{\gamma}_1}{2} + \widehat{\gamma}_2 \left(age_{i,t} + \frac{1}{4} \right) + \frac{\widehat{\widehat{\gamma}}_3}{2} \widetilde{z}_{i,t}$$

which implies that for unit i in PSU k, the approximation error is given by

$$F_i - F_k^{BLS} = \widehat{\gamma}_2 \left(\frac{1}{4} + age_{i,t} - \overline{\text{age}_{k,t}} \right) + \frac{\widehat{\widehat{\gamma}}_3}{2} \left(\widetilde{z}_{i,t} - \overline{\widetilde{\text{z}}_{k,t}} \right) \tag{5}$$

The presence of $\frac{\widehat{\gamma}_2}{4}$ term in (5) is a consequence of the infinitisemal-time approximation, and – since $\widehat{\gamma}_2$ is typically positive – implies that F_k^{BLS} is biased downwards. However, in 2004 data this coefficient is on the order of 10^{-3} or smaller. To produce a ceteris paribus comparison, we computed (4) using the unit-by-unit age-bias factors mentioned immediately above as the baseline, and using unit-by-unit age-bias factors computed using the correct non-infinitesmal formula as the alternative. Here, the error is quite small, resulting in an downward bias of less than .003% in almost every PSU, and an overall downward bias of about .001%.

The presence of the $\widehat{\gamma}_2 \left(age_{i,t} - \overline{\text{age}_{k,t}} \right)$ and $\frac{\widehat{\widehat{\gamma}}_3}{2} \left(\widetilde{z}_{i,t} - \overline{\widetilde{\text{z}}_{k,t}} \right)$ terms in (5) is a result of the PSU-average approximation. Referring back to (1), notice that this approximation introduces a source of error into the BLS rent-relative computation. In particular, this amounts to a distortion of the expenditure-based weights w_i – i.e., it incorrectly increases the *relative* importance of some units

compared to others – with units whose *age* or z is different from the PSU average receiving a weight distortion, the sign and size of which depends upon the signs and sizes of the estimated coefficients $\widehat{\gamma}_2$ and $\widehat{\gamma}_3$. Since this term is exponentiated in (1), it will *not* cancel out across units, and hence this approximation will introduce bias. (Furthermore, the bias in Rent might well be different from the bias in OER, owing to their different aggregation weights.)

How large is the error due to this approximation? We computed (4) for the 87 PSU's, for the four Census regions, and for the entire US, computing unit-by-unit factors rather than PSU-average factors, but continuing to use the infinitesimal-time approximation. The largest rent-relative adjustment for a "published" PSU was 0.99944 for Chicago, implying that this approximation error (ceterus paribus) caused inflation to be overstated in Chicago by perhaps .06%. (Conversely, inflation was understated in Phoenix by about .05%.) Overall, the bias caused by this approximation error on the US rent index was upward, but by less than .001%.

Thus, the approximation errors appear to be insignificant, and largely offsetting, if one remains in the context of (2), the BLS hedonic model. However, we argue below that the BLS specification has important weaknesses. And as noted above, a PSU-average approach in conjunction with an alternative specification – such as a piecewise-linear or higher-order polynomial in age – could impart *significant* bias upon estimated age-bias factors. Indeed, with a third-order polynomial, this approximation will *readily* generate age-bias factors of the incorrect *sign*. (This is ultimately because the effect of age is now quadratic, so that the (weighted) average effect of age can differ substantially from the effect evaluated at the (weighted) average age.) Similarly, higher-order terms in age make the infinitesimal-time approximation more questionable. Hence, in the sequel we compute all alternative age-bias factors without making these approximations.

If a common PSU-wide Rent or OER factor is required for production or reporting purposes, the weighting in (4) implies that this should not be computed as the simple average of the factors in the PSU. Instead, each of these two factors should be computed for PSU k as

$$\widetilde{F}_{k,t} = \ln\left(\frac{\sum_{i\in PSU(k)} w_i r_{i,t-6} e^{F^{ALT}_{i,t}}}{\sum_{i\in PSU(k)} w_i r_{i,t-6}}\right)$$

These factors are easily computed; note the similarity to (1). The Rent and OER factors will differ in general, since they do not share the same distribution of relative weight across age; they will be

identical only for very simple regression specifications.[11,12]

6 Specification Issues: Leverage, Disaggregation, and Model Selection

6.1 Leverage

Leverage is a key issue in age-bias estimation. Recall from Section 2.2 that obtaining unbiased estimates of the $\widehat{\gamma}$ vector itself is crucial. But in (2), when taken as a group, units aged 101 years or more – which comprise about 5% of the sample – have high leverage, on the order of twice that of the typical unit. (Units aged 201 years or more, comprising 0.2% of the sample, have about *ten* times the leverage of the typical unit.) However, even this statistic understates the influence of old units: such units form an "outlier group," in which the presence of other members in the group masks the importance of any particular individual. (Robust regression techniques such as least trimmed squares are a potential solution to this type of problem, but are too costly to implement except in relatively small data sets.)

It is easy to illustrate the negative consequences of aged units. We estimated three models using all units in the sample. The first is the official BLS specification, which – in addition to other non-age-related regressors – includes age, age^2, and three interaction term, $age_i \cdot I_i^{age>85}$ (where $I_i^{age>85} = 1$ if unit i's age is greater than 85), $age \cdot allrooms$ and $age \cdot detached$. The other two models were an age-bin model, and a three-part-spline model which had second-order terms in the first two parts, and featured knots at ages 26 and 85. (For comparability, we included $age \cdot allrooms$ and $age \cdot detached$ in these latter two models as well; estimated coefficients on these terms are very similar across the three models, and qualitative results are not sensitive to keeping them in or leaving them out.) Note that the $age - age^2$ specification (without any further age-interaction

[11] Indeed, for the entire US, the 2004 Rent weights are quite variable across groups of ages, and turn out to be largest on units aged 70-100 years, followed by units aged 18-34 and over 100. Conversely, the 2004 OER weights are less variable across groups, but largest on units aged 50-70 and on units aged over 100 years.

[12] Suppose that operational considerations require that a *single* factor be produced for each PSU. Denote OER weights by w_i, and Rent weights by v_i. If we assume that the final criterion is to minimize the weighted sum of squared errors, with weights ϕ and $(1-\phi)$ on the squared errors of OER and Rent respectively, then the PSU factor $\widetilde{F}$ should be computed as $\widetilde{F} = \ln\left(\frac{\phi \sum w_i r_i \sum w_i r_i e^{F_i} + (1-\phi)\sum v_i r_i \sum v_i r_i e^{F_i}}{\phi(\sum w_i r_i)^2 + (1-\phi)(\sum v_i r_i)^2}\right)$

terms) is common in the hedonic rent literature.[13]

As is evident in Figure 1 below, the standard BLS specification does not adequately capture the effect of age: there is a noticeable *understatement* of the rent-reducing effect of age for units aged 26 years or less (which will be consequential, since about 30% of the sample has age < 26 years), and an *overstatement* of the rent-reducing effect of age for older units. Removing units aged 100 years and older brings a marked improvement to the fit of the BLS model (these results are not depicted, so as to keep the figure uncluttered). Evidently minimizing least-squared errors with such an inflexible functional form resulted in an overemphasis upon very aged units, which comprise a trivial fraction of the sample. Conversely, a three-piece spline does a much better job fitting the true conditional age-log rent profile.

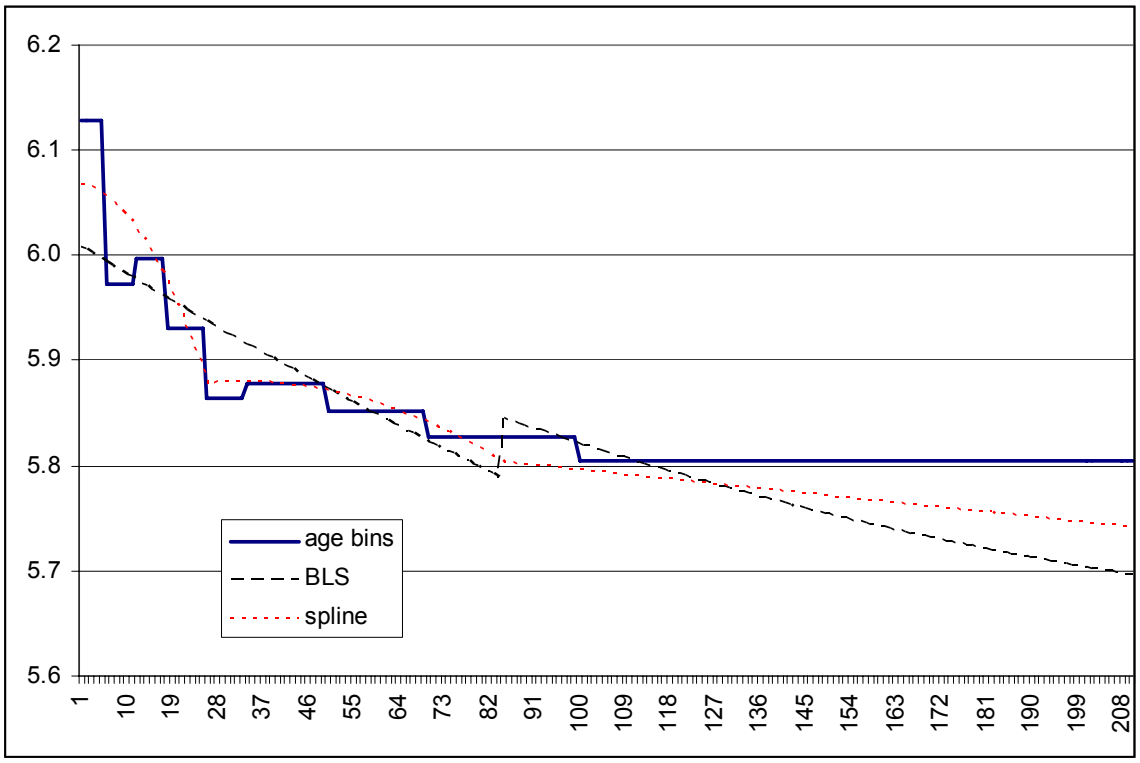

Figure 1: Leverage with age-age^2

The leverage problem may also be diagnosed by comparing coefficient estimates upon restricting

[13] See, e.g., Crone, Nakamura and Voith (2003) and Gordon and vanGoethem (2004).

the sample to units aged 200 and less, and then units aged 100 and less: the estimated coefficient on *age* triples in size, and the estimated coefficient on age^2 increases by an order of magnitude, despite a 5% change in sample size. Clearly, this will have a substantial impact on estimated age-bias factors.

It is desirable to avoid undue influence of old units on age-factor estimates, since these comprise only a small fraction of the sample. Thus, we approached the problem by splitting the sample into two parts: the 95% of the sample comprised of units aged 100 years or less, and the remainder. We treat each part of the sample separately, as described in more detail below.

6.2 Disaggregation: location, location, location

A key issue in nearly all empirical work related to real-estate is, not surprisingly, location. The current official BLS aging-bias regressions are conducted on a Census-region basis, with dummy variables indicating PSU-size. But should one thus impose equality of coefficients (including the constant) across all the cities within a Census region?

Both theory and informal evidence suggest that this is not appropriate. The real estate markets of Honolulu, Anchorage, San Francisco and Denver – all cities in the "West" – do not move in lockstep; nor is the importance of such features as air-conditioning identical across these cities. (In keeping with this, it is probably unadvisable to impose common effects of deterioration across diverse cities.) One can also formally test the non-equality of particular regression coefficients gracefully in the context of a single regression; in each of the handful of cases we investigated, formal tests of equality of particular regression coefficients between PSU's also rejected the null hypothesis of equality. As we report below, F-tests for the exclusion of PSU dummies in regional regressions strongly reject the null hypothesis, indicating that these variables should not be omitted from the regression specification,[14] and suggesting that further disaggregation might be appropriate.

But what about the danger of overfitting? This is a valid concern, but one which is partly addressed using the cross-validation procedure which we describe below; if increasing the level of disaggregation yields vastly superior *out-of-sample predictions* to those obtained from a smaller

[14] Previously, the official BLS specification included PSU dummy variables; see Campbell (2006).

amount of disaggregation, this does much to alleviate concerns related to overfitting.

However, we again remind the reader that superior predictive ability does not necessarily imply superior age bias adjustments. Unbiased age bias adjustments rely upon coefficient estimates which accurately estimate the marginal effect of age; but one simply cannot guarantee this, even if the correct level of disaggregation were known. Standard error estimates are not informative to this question. The fact that one can estimate, with great *apparent* precision, the *all-US-average* reduction in rent caused by an extra year of aging, does not imply that this estimate is an accurate estimate of the true effects of an extra year of aging on a typical unit in Sacramento. In other words, one simply cannot quantify the benefits or costs relating to the larger-but-less-relevant samples. Superior out-of-sample predictive ability is surely related to improved *overall* model accuracy, but does not necessarily mean improved correspondence of estimated coefficients to true marginal effect. For this reason, we argue below that a model averaging approach is important.

Given the importance of location, we considered a disaggregation scheme which placed PSU's into fourteen groupings, which are listed in the Appendix. In each group – including two "groups" which consist of a single PSU – there is a minimum of 200 degrees of freedom, and generally an order of magnitude more. In some of the models we considered, we also investigated the usefulness of $age \times PSU$ interaction terms. Preliminary data analysis using three different models indicated that the 14-region level of disaggregation was far superior to the four-region level in terms of out-of-sample prediction.

Having considered leverage and disaggregation, we now turn to the issue of model selection.

6.3 Model selection

6.3.1 Model uncertainty and model averaging

Empirical research typically aims to determine the degree of support for hypotheses about unknown parameters. Usually, researchers will provide information both about point estimates and about their reliability, or about multivariate analogues such as forecasts or impulse response functions (along with confidence intervals). Hedonic adjustments are often conducted on the basis of

coefficient estimates, so it is naturally desirable that these estimates be precise.

However, prior to any of this, an empirical model must be selected.

Theory should guide a regression specification. But theory uncertainty is common; i.e., there are often competing theories which try to explain economic outcomes. Thus, a researcher may have little guidance about the empirical specification, and there may be a large list of potential independent (or "control") variables.

Why is this a problem? Suppose we are interested in estimating "the marginal influence of x_i on y" – which, loosely speaking, amounts to estimating β_i, a coefficient in some appropriate regression model. But which model? Generally, the researcher does not know for certain which other variables enter into the regression – and may not have the relevant data in any case. The researcher may not know for certain the correct specification in x_i: should this variable enter as $\ln(x_i)$, as x_i, as $(1+x_i)^{\frac{1}{6}}$, or in some other manner? There is no way to guarantee that the Gauss-Markov conditions hold, and good reason to doubt that they hold when there are important omitted variables. Hence, even if estimated very precisely, there is no way to guarantee that any particular coefficient estimate is truly capturing "the marginal influence of x_i on y." Furthermore, coefficient estimates can vary substantially across models; that is, estimates can be *sensitive* to changes in specification, so that inferences can be fragile. (Indeed, this motivates pleas for sensitivity analysis, which Magnus (2006) defines as "the study of the effect of small changes in model assumptions on an estimator or test statistic of a parameter of interest.")

This model-uncertainty problem is typically ignored almost completely. In usual practice, a researcher begins by choosing a small or large set of potential models, and uses some selection criterion – such as step-down testing, information criteria, or informal specification searches guided by t-statistics – to select a single model.[15] After this, inference then proceeds as if the model is correct and as if this model selection had not taken place. But in the context of nonexperimental data, when there is fundamental uncertainty about the data generating process, presenting the results of a single preferred regression model can vastly understate the degree of uncertainty about parameter values. Indeed, this practice induces size distortions and can be dramatically misleading

[15] We further suspect that in many cases, model selection is incomplete until the results are "acceptable," i.e., are "statistically significant" and match the priors of the researcher. An advantage of the Bayesian approach to statistics is that the researcher must reveal his or her priors.

(see Freedman 1983 and Raftery 1995): standard theory is based on the assumption that the model in use was specified without any data-dependent modeling choices, whereas typical practice uses the data to reject *many* models before a particular model is chosen.[16] (Even the simple "innocuous" practice of omitting "insignificant" variables will typically lead to unjustifiable confidence in parameter estimates.) But in most studies, model uncertainty only receives a cursory glance (at best) via the presentation of regression results from several closely-related models.

Different model selection procedures account for some of the divergence in coefficient estimates that, in turn, incite bitter empirical battles in many literatures. The flip side of this is that typically *several* different models may all seem reasonable given the data, yet these models may lead to very different conclusions regarding particular parameters. Furthermore, the true data generating process is likely far more complex and subtle than *any* of the models being entertained by the researcher. Thus, any particular model must be viewed as being one approximation among many; and it seems implausible that any one empirical specification truly captures "relevant reality," or that any particular model will dominate all others at every point in the domain. Model uncertainty is thus a key issue facing almost all empirical analysis, and ignoring it does not make it go away. (See Temple (2000) for a more thorough discussion of model uncertainty.)

A recent approach to the general problem model selection and model uncertainty starts with the admission that one does not know which model is true (or, when one knows that there are important omitted variables, with the admission that *none* of the models is true), and then does what is sensible: take averages over models. This is an approach deriving from Leamer (1978)[17] which is increasingly gaining wide acceptance. Both theoretical and empirical evidence support model averaging. In a forecasting context, Makridakis and Winkler (1983) explain one aspect of this as follows (page 990): "When a single method is used, the risk of not choosing the best method can be very serious. The risk diminishes rapidly when more methods are considered and their forecasts are

[16] Raftery (1995) is a key reference, and provides a more thorough discussion of many of the points noted here. Note also that there is a large literature discussing the effects of model selection on inference. For example, Potscher (1991) shows that AIC selection results in distorted inference; and Kabaila (1995) examines the impact on confidence regions. Caudill and Holcombe (1999) explore two popular specification search methods and show that these can readily lead to spurious t-statistics. Danilov and Magnus (2004) show that ignoring model selection can generate substantial error in the prediction interval. Unit-root inferences are notoriously susceptible to changes in model-selection methods.

[17] Leamer (1978) argues that there are six distinct reasons for specification searches, which lead to six different varieties of search procedures.

averaged." There is a large Bayesian literature (see Hoeting, Madigan, Raftery, and Volinsky 1999 for a survey), and a growing frequentist literature (see Buckland, Burnham and Augustin 1997; see also Magnus and Durbin, 1999). In economics, model averaging has become commonplace in the cross-country growth literature[18] – and is increasingly dominant in the forecasting literature. In particular, since Bates and Granger (1969), a large body of research in the forecasting literature has confirmed that combinations of individual forecasts often outperform individual model forecasts, in the sense that the combined forecast delivers a smaller mean-squared forecast error (MSFE); see, e.g., Stock & Watson (2004).[19]

Model averaging is straightforward. To implement this technique in the conventional manner, initially L models are selected, where the set of models might have already been reduced via a model selection procedure (to eliminate clearly inferior models).[20] Then estimates are formed by weighted averages. For example, to form the prediction $\widehat{y}$, one forms a weighted average over the predictions of the L models:

$$\widehat{y} = \sum_{l=1}^{L} w_l \widehat{y}_l.$$

As long as one is using weights which have been fixed beforehand and which sum to one, then if each individual model yields predictions which are unbiased, this weighted-average prediction will also be unbiased.

Estimating a parameter θ (assumed common to all models) is accomplished in the same way;

[18] In the economics literature, six recent Bayesian-model-averaging studies are Fernandez, Ley and Steel (2001), Brock and Durlauf (2001), Koop and Potter (2003), Sala-i-Martin, Doppelhoffer and Miller (2004), Eklund and Karlsson (2005), and Masanjala and Papageorgiou (2007). See also Brock, Durlauf and West (2003) for insightful comments, and Durlauf, Johnson and Temple (2005) for technical advice.

[19] By combining forecasts from several models, the forecaster implicitly acknowledges that more than one model could provide good forecasts, and guards against misspecification, poor estimation, or instabilities/non-stationarities by not putting all the weight on one single model (see Hendry and Clements 2004, and Timmermann 2005). Furthermore, it can be shown (see Timmermann 2005) that even if the forecasts from one model dominate those from another model (in the sense that they lead to lower expected loss), it may still be optimal to combine the two forecasts.

[20] Swanson and Teng (2001) propose using a criterion like this to choose which subset of forecasts to combine. Others use such criteria as the basis of weights; see below.

one forms a weighted average of $\widehat{\theta}$ over the L models:[21]

$$\widehat{\theta} = \sum_{l=1}^{L} w_l \widehat{\theta}_l. \tag{6}$$

One can also determine the variance of $\widehat{\theta}$; see Buckland, Burnham and Augustin (1997) and Sala-i-Martin, Doppelhoffer, and Miller (2004). The argument for model-averaging is even more compelling in this case than in the prediction case. As noted above, the idea is to avoid an error stemming from the use of an incorrect model. In the case of coefficient estimates, we cannot guarantee that the coefficient estimate from any particular model l, $\widehat{\beta}_l$, accurately estimates the true marginal effects of the variables in question. If these coefficient estimates vary across models which are otherwise roughly comparable in their ability to approximate the data, this is cause for concern, since any particular model is but one approximation of reality. When there are multiple reliable signals, it makes sense to average, even if they are correlated; the *average* estimate is likely to be closer to the truth than any one taken individually.

Obviously, a key practical issue is how one should determine the weights w_l. There are several approaches, one being simple averaging (i.e. setting $w_l = \frac{1}{L}$), which in the forecasting context is often difficult to beat (see, e.g., Clemen 1989 and Stock and Watson 2001). Weights might also be estimated by regression, i.e. by choosing weights to minimize the mean squared forecast error of the averaged model. But estimation errors that contaminate the combination weights are known to be a serious problem for many combination techniques; see Diebold and Pauly (1990), Elliott (2004), Hendry and Clements (2004), Yang (2004), and Timmermann (2005).

Alternatively, to construct weights, a common suggestion is to use some weighting criterion

[21] Suppose one is particularly interested in a particular coefficient estimate. Since a particular variable x_k might not appear in every model, the sum of the weights applied to the coefficient β_k will not equal unity, which will "bias" the estimate of that coefficient toward zero. An alternative construction uses *rescaled* weights, i.e. sets

$$w_l = \frac{C_l}{\sum_{i \in N(k)} C_i}$$

where $N(k)$ is the set of models which contain x_k.

But one must be careful in interpreting averaged coefficients. A coefficient estimate in a particular regression model equals, at best, the marginal influence of its regressor *conditional on the presence of all the other variables in the model.* Thus, there is no reason to suspect that the coefficient estimates from two different specifications should be equal: *they are estimates of two conceptually different things.*

C_l,[22] and apply it in a formula such as

$$w_l = \frac{C_l}{\sum C_i}.$$

In a Bayesian context, the weighting criterion is the posterior probability. However, a purely Bayesian approach is rarely used; see Jacobson and Karlsson (2006), and the discussion in Shtatland et al. (2000) and Yuan and Yang (2005). A commonly-used approximation to the Bayes factor is the Bayes-Schwarz information criteria; see Raftery (1995).[23]

An alternative approach is to use out-of-sample cross-validation methods. Cross-validation, due to Allen (1974), is a commonly used model selection criteria, with various consistency results; see, e.g., Yang (2005). In general terms, the data is split into two parts: $N - k$ observations, to be used for fitting each competing model (or procedure), and the remaining k observations, to be used to measure the performance of the models. A common performance measure is the MSFE on the k reserved (out-of-sample) observations. In a cross-section context, it is straightforward to iterate upon this procedure, either via partitioning the sample into n equal-sized parts (with k observations in each), or by randomly selecting the k observations each iteration. Such methods base model selection, or weights in model averaging, either upon appropriate ratios of MSFEs, or upon the fraction of iterations a particular model wins the implicit horse race. (See Pesaran and Timmerman 2006, who – in a context of forecasting under uncertainty about break dates – compare

[22] The criterion might be the inverse of mean squared error, or an exponentiated information criterion IC_l, i.e., $C_l = \exp\left(-\frac{1}{2}IC_l\right)$.

[23] Shtatland et al. (2000) note that the AIC and SIC can emulate the Bayesian approach under two opposite and situations. Model comparisons based on AIC are asymptotically equivalent (see Kass and Raftery, 1995) to those based on Bayes factors, under the assumption that the precision of the priors is comparable to that of the likelihood (in other words, only if the information in the prior increases at the same rate as the information in the likelihood, so that the prior is as informative as the data). Conversely, $exp(-\frac{1}{2}SIC)$ provides a surprisingly good approximation to the Bayes factor (see Kass and Wassermam, 1995) when the amount of information in the prior is equal to that in one observation (at least when comparing nested models), so that the prior is not informative at all. Shtatland et al. (2000) recommend following a standard model selection procedure (such as step-down testing), then determining the model favored by AIC, the model favored by BIC, and all models "in between," i.e. those which lead (by one's selection process) from the larger AIC-favored model to the smaller BIC-favored model. They term this set of models (from AIC through BIC) the "AIC-BIC window," and recommend averaging over these models using one of the criterion. As these researchers point out, this model selection procedure is straightforward to implement and avoids the conceptual and computational difficulties associated with a purely Bayesian approach. Of course, their method does not generalize to situations in which models are inherently non-nested. Hansen (2006) provides evidence that selecting weights by minimizing a Mallow's criterion, which is an estimate of the squared error, is superior to using exponentiated-AIC or BIC weights. The focus of the investigation might determine the appropriate criterion. In the context of obtaining a reliable coefficient estimate (as opposed to a reliable forecast), the risk associated with including an irrelevant variable is lower than the risk of excluding a relevant variable, which would favor AIC over BIC.

equal weights to weights which are chosen to be proportional to the inverse of the MSFE values.) One key advantage of the implicit-horse-race method is that the resultant weights are not distorted by the presence of a large number of similar models.

Granger & Jeon (2004) suggest a thick-modeling approach, based on trimming to eliminate the $k\%$ worst performing forecasts, and then taking a simple average of the remaining forecasts; this concurs with a conclusion of Hendry and Clements (2004), who state "since otherwise, one really poor forecast would worsen the combination needlessly."

In this study, we use a combination of cross-validation and thick-modeling, as described below.

6.3.2 Applying model averaging to age-bias estimation

In the age-bias context, the true data-generating process is almost certainly more complicated than any estimable model, if only because – in this real-estate context, where the rule of thumb is "location-location-location" – a large number of neighborhood variables are missing. Furthermore, coefficient estimates turn out to vary across different specifications. Thus, in the age-bias context, the argument for model averaging is quite compelling.

How would one implement such averaging? Once weights are chosen, it is straightforward to apply averaging to age-bias estimates from different models – i.e., for each unit i, to construct

$$\widehat{age\text{-}bias}_i = \sum_{l=1}^{L} w_l \cdot \widehat{age\text{-}bias}_{i,l} \tag{7}$$

Which criterion should be used in selecting weights? If the goal were to impute missing log-rent observations, a pure cross-validation approach would be the natural choice. However, the goal in age-bias estimation is to obtain a reliable estimate of the effect of increased age on log-rent, which is not necessarily guaranteed by a model which reliably predicts log-rent out-of-sample. There are two reasons for this. First, as noted above, there is no guarantee that individual coefficient estimates accurately capture the true marginal effect of age. Second, and somewhat surprisingly, even cross-validation could lead to overfitting in the present context. This is due to the sampling procedures underlying our data. The geographic stratification scheme employed by the BLS begins by dividing a PSU into six regions, and then into geographic "segments" in each region, which

are Census blocks or block-groups. Segments from each region are randomly selected proportional to their "size" or shelter expenditure. Once a segment is selected, the goal is to obtain five or more rental housing units from each segment. But the rental housing in a Census block is often similar along many characteristics, including both log-rent and age. If this is the case, even a cross-validation procedure might lead to "overfitting" the data along dimensions of age, with age proxying for the missing neighborhood variable (and thereby helping the model to predict missing observations). This could occur even though a large number of neighborhood variables, such as percent-renter, are included in the regression.

For these reasons, in this study we selected weights using a combination of cross-validation, trimming, and simple averaging, as follows. First, we conducted an extensive cross-validation exercise as an initial "filter," examining a number of alternative specifications at various levels of disaggregation, to weed out poorly-performing models. (We report some of the findings of this study below.) We conducted several iterations of the cross-validation exercise, as some specifications were refined based upon the results of previous iterations. Upon obtaining seven reasonable models, we then followed Granger and Jeon (2004) and, on a unit-by-unit basis, trimmed the highest and lowest estimated age-bias estimates; our age-bias estimate is the simple average of the remaining estimates. (Our decision to trim estimates and average in this way was partly motivated by the fact that, as we discuss below, one of our best-predicting models sometimes generated implausible age-bias estimates. Note that one could use trimming in combination with weighting; simply include multiple copies of each estimate in the pool of estimates to take a trimmed mean over, tying the number of copies of an estimate to the relative weight on that model.)

Is this procedure infeasible? Most other statistical agencies in the world simply lack the man-power to undertake a study similar to this one. Furthermore, many statistical agencies might resist the idea of model averaging in any case, since it is difficult to motivate and explain to the public, and thus could end up having a bit of a "black box" character. Our response is threefold. First, statistical agencies still must perform model selection in any case; the amount of additional effort required to implement a simple form of model averaging is minimal. In particular, instead of discarding all but one model, analysts could retain several of the top candidates, specify simple-average weights or information-criterion weights, estimate each model, and form (7) as above. Second, the

risk to using the wrong model can be substantial. Third, it is possible to describe this procedure to the public in a simple manner. Here are two possibilities. First, it could be described as an *estimation* technique: "Given our desired large model, we estimate each coefficient estimate in this large model using (6)." Second, model averaging over the l surviving candidate models could be *described* as estimating a single large model as in

$$\ln rent_i = \frac{1}{l} f_1\left(age, \widetilde{X}; \widetilde{\alpha}_1\right) + \frac{1}{l} f_2\left(age, \widetilde{X}; \widetilde{\alpha}_2\right) + ... + \frac{1}{l} f_l\left(age, \widetilde{X}; \widetilde{\alpha}_l\right) + u_i$$

where $u_i := \frac{1}{l}\left(u_{1i} + u_{2i} + ... + u_{li}\right)$, $\widetilde{\alpha}_j$ is the coefficient vector corresponding to model j, and each model specifies *age* and/or various elements of $\widetilde{X}$ differently. Of course, estimation must still *proceed* via estimating each model j separately. (The system should not be estimated jointly; although this could be readily accomplished upon dividing each element in each model by $\frac{1}{l}$, it could run into degrees-of-freedom difficulties. Even if it didn't, joint estimation would still likely lead to overfitting the data.)

6.3.3 Candidate models

There are obviously an enormous number of alternative model specifications one could consider. In practice, one must rely upon intuition to help narrow the search to a manageable number. Given the preponderance of hedonic studies using log-price as the dependent variable, we did not examine any alternative, and used log-rent as our dependent variable in all cases.[24] Given our findings regarding leverage, we split the sample into two parts: the 95% of the sample comprised of units aged 100 years or less, and the remainder. We modeled each part of the sample separately, except as noted below.

For units with $age \leq 100$, we considered ten alternative models.[25] Preliminary work indicated the necessity including PSU dummy variables, so these are included except as otherwise noted. Barring any strong evidence suggesting the contrary, for each model $\widetilde{X}$ includes the unit-level characteristics (such as number of rooms, and whether the structure is detached or multi-unit) and the Census-2000 neighborhood variables (such as % of population that is white, and % of

[24] Since the index is moved by a rent relative, the standard log-bias adjustment term cancels out.

[25] Some key omitted models are those which are estimated using panel methods. There are several advantages to panel estimation; for example, this allows one to control for unit-specific effects. As results are so starkly different, we explore this issue in Gallin and Verbrugge (2007).

population that is under the poverty line) which are in the official BLS aging-bias specification. (The complete list of variables is given in the appendix.) In principle, information criteria such as AIC could be used to help determinine whether "marginally significant" regressors should be included in any particular regional specification;[26] however, we do not make systematic use of such criterion in this paper, except for answering broad questions such as "should PSU dummy variables be included?" The models derive from the following six specifications:

- Standard BLS model with PSU dummy variables: within group j,

$$\ln rent_i = \alpha + \widetilde{\beta}\widetilde{X_i} + \gamma_1 age_i + \gamma_2 age_i^2 + \widetilde{\gamma}_3 \widetilde{z_i} age_i + \sum_{PSU(k)\in j} \gamma_{4,k} I_i^{PSU(k)} + u_i$$

  where $I_i^{PSU(k)} = 1$ if unit i is in $PSU\,(k)$, and 0 otherwise. We examined this model at the 14-group level, and – for comparison purposes – the standard BLS model, i.e., the model without PSU dummy variables, estimated on the full sample, at the four-Census-region level of disaggregation.

- Augmented $age - age^2$ model: within group j,

$$\ln rent_i = \alpha + \widetilde{\beta}\widetilde{X_i} + \gamma_1 age_i + \gamma_2 age_i^2 + \widetilde{\gamma'}_3 \widetilde{z_i'} age_i + \sum_{PSU(k)\in j} \gamma_{4,k} I_i^{PSU(k)} + \sum_{PSU(k)\in j} \gamma_{5,k} age_i I_i^{PSU(k)} + u_i$$

  This model augments the standard BLS model by including PSU dummy variables and more age interaction terms, including $age \times PSU$ interaction terms. This model was examined at the the four-Census-region level of disaggregation, as well as at the 14-group level. In this case and in many of the cases following, we used informal criteria – such as insignificant t-statistics – to eliminate particular $age \times PSU$ or $age \times z$ terms.

- Piecewise-linear: within group j,

$$\ln rent_i = \alpha + \widetilde{\beta}\widetilde{X_i} + \gamma_1 age_i + \gamma_2 I_i^{age\geq s}\,(age_i - s) + \widetilde{\gamma}_3 \widetilde{z_i} age_i + \sum_{PSU(k)\in j} \gamma_{4,k} I_i^{PSU(k)} + u_i$$

  where s is the knot, i.e. the point at which the first linear piece intersects the second linear piece. (Note that this specification imposes the restriction that this intersection occurs at s, i.e. the piecewise-linear fit to the data is continuous.) Here, the knot was chosen to be age 26, i.e., $s = 26$, based upon AIC.

- Two-piece and three-piece spline: within group j,

$$\begin{aligned} \ln rent_i \;=\; & I_i^{age<s_1}\left(\alpha_1 + \delta_1 age_i + \theta_1 age_i^2\right) + I_i^{s_1\leq age<s_2}\left(\alpha_2 + \delta_2 age_i + \theta_2 age_i^2\right) \\ & + I_i^{age>s_2}\left(\alpha_3 + \delta_3 age_i\right) + \widetilde{\beta}\widetilde{X_i} + \widetilde{\gamma}_3 \widetilde{z_i} age_i + \sum_{PSU(k)\in j} \gamma_{4,k} I_i^{PSU(k)} + u_i \end{aligned}$$

[26] AIC is perhaps preferable to BSIC in this context, given the danger of omitted variable bias. See Shtatland et al. (2000), who suggest combining stepwise regression techniques with information criteria in order to avoid the intractible problem of searching a combinatorial number of models.

where, in the two-piece case, $s_1 = 0$ and $s_1 = 50$, and in the three-piece case, $s_1 = 26$ and $s_2 = 80$, and where the restrictions $\alpha_1 + \delta_1 s_1 + \theta_1 s_1^2 = \alpha_2 + \delta_2 s_1 + \theta_2 s_1^2$ and $\alpha_2 + \delta_2 s_2 + \theta_2 s_2^2 = \alpha_3 + \delta_3 s_2$ must be imposed. These models were estimated on the full sample, rather than only on units aged 100 years and less. In some cases, higher-order terms were also considered.

- Chebyshev polynomial: within group j,

$$\ln rent_i = \alpha + \widetilde{\beta}\widetilde{X_i} + \sum_{r=1}^{m} \lambda_r z_r + \widetilde{\gamma}_3 \widetilde{z_i} age_i + \sum_{PSU(k)\in j} \gamma_{4,k} I_i^{PSU(k)} + \sum_{PSU(k)\in j} \gamma_{5,k} age_i I_i^{PSU(k)} + u_i$$

  where $z := \left(\frac{age_i - 100}{100}\right)$, with higher-order Chebyshev terms z_r defined analogously, and where m is chosen between 3 and 5 (with an attempt to avoid overfitting). We considered two Chebyshev polynomial models, one a greatly restricted version of the other, with smaller m and fewer regressors.

- Log-age: within group j,

$$\ln rent_i = \alpha + \widetilde{\beta}\widetilde{X_i} + \gamma_1 \ln age_i + \widetilde{\gamma}_3 \widetilde{z_i} age_i + \sum_{PSU(k)\in j} \gamma_{4,k} I_i^{PSU(k)} + u_i$$

For units with $age > 100$ (which comprise about 5% of the sample), we considered the two 14-region spline models described above (as before, estimated on all units), the standard BLS model at the four-Census-region level of aggregation (estimated on all units), and two other models at the all-US level of aggregation. The first of these is very simple, having only one term in age, namely $\ln age$, 30 other regressors (including a constant, 11 PSU dummy variables, and one region dummy variable), and is estimated only on units with $age > 100$. The other model contained only linear terms in age (including age-interaction terms), and was estimated on all units with age between 40 and 300. Based upon a small cross-validation study, our final age-bias factor for these older units is the average of the factors from the latter two models and the three-part spline model.

Each model likely suffers from heteroskedasticity; but this is not a problem because we do not rely on standard error estimates, and heteroskedasticity does not bias coefficient estimates. Outliers are likely not problematic, for two reasons. First, sample sizes are generally comfortably large. Second, the effect of outliers on model selection is muted by our procedure: we use a cross-validation horse-race procedure, which will penalize both overfitting and excessive sensitivity to outliers. (One might still wish to consider the possible effects of outliers on final model estimation, once model selection and averaging weights have been determined.)

In our cross-validation procedure, for each iteration, we reserve 10% of the sample for testing forecasts. Each of the models is fit (on each of the 14 groups) on the fitting portion of the data, and then each model is used to forecast the remaining 10% of the observations in that group. At the end of the iteration, the overall MSFE – summed across all units – is computed. Then the procedure is repeated.

6.3.4 Cross-validation results

The results of the cross-validation study for units aged 100 years and less are given below. (We conducted a similar study on older units, and also examined the performance on a region-by-region basis, but do not report these results in the interest of brevity.)

Define

$$\begin{aligned} \overline{dev}^k &: = \frac{100}{N} \sum_{s=1}^{N} \frac{MSFE_s^k - MSFE_s^{best}}{MSFE_s^{best}} \\ mdev^k &: = 100 \cdot \underset{s}{Max} \frac{MSFE_s^k - MSFE_s^{best}}{MSFE_s^{best}} \end{aligned}$$

where $MSFE_s^k$ is the MSFE for model k in iteration s, and $MSFE_s^{best}$ is the smallest MSFE (across models) for iteration s. The first measure, $\overline{dev}^k$, is the average percentage increase in MSFE corresponding to model k; for example, if $\overline{dev}^k = 4\%$, then the use of model k implies an MSFE that is, on average across the N iterations, 4% higher than the best model. Similarly, the second measure, $mdev^k$, is the *maximum* observed percentage loss in MSFE corresponding to model k. We also report the overall MSFE win percentage for each model k, i.e., the percentage of iterations for which $MSFE_s^{best} = MSFE_s^k$.

Based upon 500 iterations, results are as follows:

	BLS	BLS +PSU	Aug. -4	Aug. -14	Piec. Lin.	2- Spl.	3- Spl.	Cheb.	Cheb. (rest.)	Log
Win %	0%	0%	9%	1%	0%	5%	1%	4%	79%	0%
$\overline{dev}$	45%	3.6%	2.2%	1.5%	2.7%	1.6%	2.1%	1.4%	0.2%	2.4%
$mdev$	56%	10%	8%	9%	10%	9%	10%	8%	7%	10%

Table 2

In Table 2, "BLS" refers to the official specification based upon 4-regions, "BLS +PSU" refers to the official specification plus PSU dummy variables (estimated on age<100), "Aug. -4" and "Aug. -14" refer to the Augmented $age - age^2$ model (estimated on 4 and 14 regions, respectively), "Piec. Lin." refers to the Piecewise Linear specification, "2-Spl." and "3-Spl." refer to the spline models (estimated on all ages), "Cheb." and "Cheb. (rest.)" refer to the Chebyshev and restricted-Chebyshev specification, and "Log" refers to the Log-age specification.

In terms of MSFE, the best model overall is clearly the restricted Chebyshev model. But surprisingly, a carefully-specified augmented $age - age^2$ model, estimated at the highly-aggregated four-Census-region level (on $age < 100$), is actually competitive with more disaggregated models, despite its multitude of implied coefficient restrictions.[27] This suggests that the 14-region level of disaggregation may be too aggressive. A common feature of these two best-performing models is the presence of the $age \cdot I^{PSU(k)}$ interaction terms, which underscores the importance of adequate treatment of location. Notice that the only model which is clearly rejected out-of-hand is the official BLS specification; even the 14-region $age - age^2$ specification, which adds only PSU dummies to the standard BLS specification (and is estimated on $age < 100$) is, on average, a mere 3.6% worse than the best-predicting model in terms of MSFE (although across regions, there are five regions in which it is outperformed in the 5-8% range).

[27] This model contains numerous $age \cdot I^{PSU(k)}$ interaction terms; F-tests for their inclusion have p-values less than 0.0000. Across the fourteen regions, the MSFE evidence is a bit more mixed. This four-region model outperforms all the other models in three of the fourteen regions, and is within 4% of the best in five others. But on the other hand, in four of the regions, it is outperformed by 9% or more.

These results also suggest that the 2-spline is a strong model. However, this model generated age-bias estimates which were clearly outside the bounds of possibility for most units in some regions (despite very good predictions on those regions). In other words, despite its superior out-of-sample prediction performance, we concluded that this model was probably overfitting the data – a possibility which is discussed in Section 6.2.2 – and we eliminated this model from consideration.

6.4 Constructing our final age-bias estimates

For units aged 100 years and less, we constructed age-bias estimates as follows. On the basis of the cross-validation results, we chose seven models – the four- and fourteen-region augmented $age-age^2$ models, the fourteen-region $age-age^2$ model (with PSU dummies), the restricted Chebyshev model, the three-part spline, the log-age model, and the piecewise-linear model. Using the unit-by-unit age-bias factor estimates from these models, we followed the trimmed-mean approach discussed above, trimming the highest and lowest of these, and constructing the simple average of the remainder; this became our age-bias estimate for that unit.

For units aged greater than 100 years, we constructed age-bias estimates (on a unit-by-unit basis) by taking the simple average of the age-bias estimates derived from the two all-US models and the three-part spline.

7 Results

Above, we pointed out a number of potential deficiencies in the current BLS approach to age-bias estimation. How significant are these deficiencies *in toto*?

Figure 2 below plots our estimated inflation adjustments from (4), across the PSU's whose shelter inflation indexes are published by the BLS.

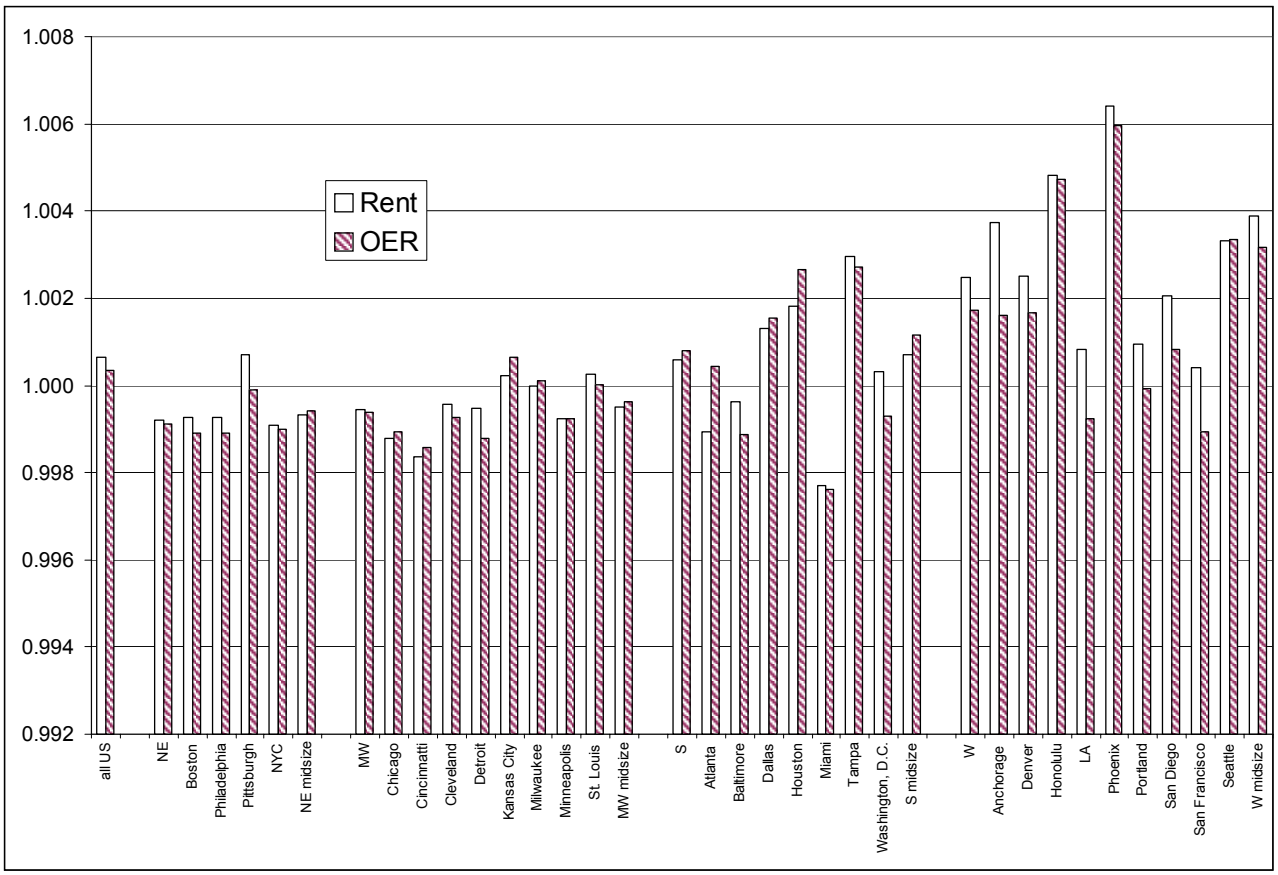

Figure 2: Estimated Adjustments to CPI Shelter Components

The average annual age-bias factor estimated by the BLS method was -0.00201; our estimated age-bias factor has an average of -0.00278, which is about 38% larger. Due to the presence of aggregation weights, this does not translate immediately into differences in rent relatives. We find that over the 2004-2005 period, the aggregate BLS Rent and OER indexes were downward-biased, but only by a small amount. According to our estimates and using (4), aggregate Rent inflation between 2004 and 2005 was downward-biased by 0.06%, and aggregate OER inflation was downward-biased by 0.04%. However, our suggested improvements can make a much bigger difference to estimates of shelter price inflation experienced by specific PSU's; the adjustments to the relative are in the range $[-0.2\%, +0.6\%]$; in other words, inflation may be overstated by as much as 0.2%, or understated by as much as 0.6%. A striking example is the country's fifth largest city, Phoenix: between 2004 and 2005, reported Rent inflation in this PSU was 0.6%, but our estimates imply that Rent inflation was double this, at 1.2%. Since OER inflation would have risen

a comparable amount as well, the use of our factors would have raised overall estimated inflation in Phoenix by about 0.2%.

Averaging over models makes a difference. For example, if one replaces our preferred age-bias factors with those from the four-region augmented $age - age^2$ model (but retaining the same age-bias factor on old units), the range of rent-relative-adjustments across PSU's is wider, namely $[-0.7\%, +0.7\%]$. These adjustments also differ, on average in absolute terms, from those derived from our averaged factors by 0.001 (with a maximum divergence as large as 0.005). (Having said this, their averages are essentially identical, and their correlation across PSU's is 0.89.)

8 Conclusion

We investigated the BLS age-bias imputation, and discovered a number of potential improvements. Do they matter? We found that, in 2004 data, BLS age-bias factors were too small, on average, by almost 40%. Since the age bias term itself is rather small, this bias had a rather modest impact on overall aggregate indexes. On the other hand, errors in particular metropolitan areas were much larger, with downward-bias in the area's shelter inflation as large as 0.6%.

We found that errors from formula approximations underlying official estimates were, in the context of the BLS model, of little consequence. However, we found more serious deficiencies related to model misspecification. In particular, the BLS hedonic regression specification – which is more or less of the form commonly used in the hedonic rent literature – is severely deficient in its ability to match the conditional log-rent vs. age relationship in the data, and performs poorly in out-of-sample tests. We found many models which are superior.

A related problem is that aging bias adjustment inherently suffers from a general problem facing some types of hedonic-based adjustments, namely the inherent impossibility of ensuring that coefficient estimates accurately estimate the true marginal impact. We advocated the use of model averaging to address this problem. This is a method that minimizes downside risks related to model misspecification and generates more reliable coefficient estimates. Simple versions of this method are easy to implement with very little additional effort.

After selecting seven "best" models using a cross-validation approach, we estimated age-bias factors by taking a trimmed average over the factors derived from each model.

We cannot argue that our estimated age-bias factors are perfect; proving or disproving this would be impossible. What we argue here is that our estimates are likely to be much better than those resulting from current BLS methods.

Currently, as a result of this study, the BLS is investigating the use of model averaging and a richer set of regression specifications. These are based upon a differential treatment of old units, including higher-order terms in age and additional age interaction terms, and an increased level of disaggregation, in conjunction with the unit-by-unit age-bias estimation that these changes necessitate.

9 Appendix

9.1 14 PSU groups

1. Northeast and and Suburbs: Boston, Hartford CT, New York City, Philadelphia, Reading PA

2. Off-Seabord Northeast Mid-Sized: Burlington VT, Johnstown PA, Sharon PA, Springfield MA

3. Mid-Atlantic Seaboard: Baltimore, Washington DC, Norfolk VA

4. "Midwest" Larger Cities: Buffalo, Chicago, Detroit, Cincinnati, Cleveland, Columbus, Dayton, Kansas City, Milwaukee, Minneapolis, Pittsburgh, Saint Louis, Syracuse

5. Midwest Smaller Cities: Brookings SD, Chanute KS, Decatur IL, Elkhart IN, Evansville IN, Faribault MN, Lincoln NE, Madison WI, Mt. Vernon IL, Saginaw MI, Wausau WI, Youngstown OH

6. Texas: Amarillo, Beaumont-Port Arthur, Brownsville-Harlingen-San Benito, Dallas, Midland, Houston, San Antonio

7. Florida: Arcadia, Gainesville, Ft. Myers, Melbourne, Miami, Ocala, Tampa

8. South Big/Medium: Atlanta, Birmingham AL, Raleigh NC, Richmond VA

9. South Small: Albany GA, Baton Rouge, Chattanooga TN, Florence AL, Florence SC, Greensville SC, Lafayette LA, Morristown TN, Oklahoma City, Picayune MS, Pine Bluff AR, Statesboro GA

10. Big West Coast: Los Angeles, Portland, San Diego, San Francisco, Seattle

11. Honolulu

12. Anchorage

13. Bigger Mountain/Desert: Denver, Las Vegas, Phoenix

14. Small West: Bend OR, Boise City ID, Chico CA, Modesto CA, Provo UT, Pullman WA, Yuma AZ

9.2 Set of regressors

There are six categories of potential conditioning variables: Census neighborhood characteristics, Unit-specific, PSU, Services-included-with-rent, Structure type, and Age-related. However, no model includes all the variables. Almost all models included all Census variables, all unit-specific variables, and the full set of PSU variables. In some regions, various services-included-with-rent were not included. Only rarely were more than one or two structure type variables included. Finally, age-interaction terms were used somewhat sparingly.

- Census neighborhood characteristic variables: % white; % in large buildings; % in mobile homes; % with 2 or more autos; % of children aged 6-18; % aged 65+; % with some college education; % lacking plumbing; % under poverty; % renter.
- Unit-specific variables: bathrooms, bathrooms^2, bedrooms, bedrooms^2, other rooms, other rooms^2, heat (electric, gas, other), air conditioning (central, window, other).
- PSU variables: Large- or Medium-sized city; PSU-dummy variables.
- Services-included-with-rent variables: heat included; electricity included; parking included.
- Structure type: detached, duplex, multi-unit with elevator, multi-unit without elevator, mobile home, other.
- Age-related: age, age^2, etc.; and age-interaction terms, of which the most common are: detached, multi-unit with elevator, $I^{age>85}$, rooms, electricity included, % white, PSU.

10 References

Allen, D.M. (1974) "The relationship between variable selection and data augmentation and a method for prediction." Technometrics, 16, 125-127.

Bates, J., and C. Granger (1969) "The Combination of Forecasts," Operational Research Quarterly, 20, 451–468.

Brock, William, and Steven Durlauf (2001), "Growth Empirics and Reality," World Bank Economic Review, 15.2, 229-272.

Brock, William, Steven N. Durlauf and Kenneth West (2003), "Policy Evaluation in Uncertain Economic Environments," Brookings Papers on Economic Activity, 235–322.

Buckland, S.T., K.P. Burnham and N.H. Augustin (1997) "Model selection: an integral part of inference." Biometrics 53, 603-618.

Bureau of Labor Statistics (2006) "Age Bias Factors Derivation Based on Unit Characteristics: 01/26/06" Internal Memo, Bureau of Labor Statistics.

Caudill, Steven B., and Randall G Holcombe (1999) "Specification search and levels of significance in econometric models." Eastern Economic Journal 25.3 (Summer), 289-300.

Campbell, Louise L. (2006) "Updating the Housing Age-Bias Regression Model in the Consumer Price Index." CPI Detailed Report, November 2006.

Clemen, R.T. (1989) "Combining forecasts: A Review and Annotated Bibliography." International Journal of Forecasting, 5, 559-583.

Crone, Theodore M., Leonard I. Nakamura, and Richard Voith (2003) "Regression-based estimates of rental increases." Mimeo, Federal Reserve Bank of Philadelphia.

Crone, Theodore, Leonard Nakamura and Richard Voith. (2006) "Rents have been rising, not falling, in the postwar period." Manuscript, Federal Reserve Bank of Philadelphia.

Danilov, Dmitry, and Jan R. Magnus (2004) "Forecast accuracy after pretesting with an application to the stock market." Journal of Forecasting 23, 251-274.

Diebold, Francis X. and P. Pauly (1990) "The use of prior information in forecast combination", International Journal of Forecasting 6, 503-508.

Draper, D. (1995) "Assessment and propagation of model uncertainty." Journal of the Royal Statistical Society, Series B 57, 45-70.

Durlauf, Steven N., P. Johnson and J. Temple. (2005). "Growth Econometrics," in Handbook of Economic Growth, P. Aghion and S.N. Durlauf, eds., North Holland, Amsterdam.

Eklund, Jana, and Sune Karlsson (2005) "Forecast combination and model averaging using predictive measures." Mimeo, Stockholm School of Economics.

Elliott, Graham (2004) "Forecast combination with many forecasts," Mimeo, Department of Economics, University of California, San Diego.

Erickson, Timothy, and Ariel Pakes (2007) "An Experimental Component Index for the CPI: From Annual Computer Data to Monthly Data on Other Goods." Mimeo, Harvard University.

Fernandez, C., Eduardo Ley and Mark F.J. Steel (2001) "Model uncertainty in cross-country growth regressions." Journal of Applied Econometrics 16, 563-576.

Freedman, David A. (1983) "A Note on Screening Regression Equations." The American Statistician 37.2, 152-55.

Granger, C. W. J. , and Y. Jeon (2004), "Thick modeling," Economic Modelling 21, 323–343.

Gordon, Robert J., and Todd vanGoethem (2004) "A century of downward bias in the most important component of the CPI: The case of rental shelter, 1914-2003. Mimeo, Northwestern University.

Hansen, Bruce (2006) "Least squares model averaging." Mimeo, University of Wisconsin.

Hendry, David F., and Michael P. Clements (2004) "Pooling of forecasts," Econometrics Journal, Royal Economic Society, vol. 7(1), pages 1-31.

Hendry, David F., and J. James Reade (2005) "Problems in Model Averaging with Dummy Variables." Mimeo, Oxford University.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial (with Discussion)," Statistical Science, 15, pp. 193-195.

Hulten, Charles R., and Frank C.Wykoff (1981) "The Measurement of Economic Depreciation." in Charles R. Hulten, Ed., *Depreciation, Inflation, and the Taxation of Income from Capital.* Washington DC: The Urban Institute Press.

Jacobson, T., and S. Karlsson (2006) "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," forthcoming, Journal of Forecasting.

Kabaila, P. (1995) "The effect of model selection on confidence regions and prediction regions," Econometric Theory 11, 537-549.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90, 773-795.

Kass, R. E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association, 90, 928-934.

Koop, Gary, and Simon Potter (2003) "Forecasting in large macroeconomic panels using Bayesian Model Averaging." Staff Report 163, Federal Reserve Bank of New York.

Magnus, Jan (2006) "Local sensitivity in econometrics," in *Measurement in Economics*, M.J. Boumans, ed., Academic Press, San Diego.

Magnus, Jan R., and J. Durbin (1999) "Estimation of regression coefficients of interest when other regression coefficients are of no interest." Econometrica 67, 639-643.

Makridakis, S. and R.L. Winkler (1983), "Averages of forecasts: Some empirical results," Management Science 29:987-996.

Masanjala, Winford H., and Chris Papageorgiou (2007) "Initial Conditions and Post-War Growth in Sub-Saharan Africa." Manuscript, IMF.

Pesaran, Hashem and Allan Timmerman (2006) "Selection of Estimation Window in the Presence of Breaks." Forthcoming, Journal of Econometrics.

Poole, Robert, Frank Ptacek, and Randal Verbrugge (2005) "Treatment of Owner-Occupied Housing in the CPI." Manuscript prepared for FESAC, Bureau of Labor Statistics.

Potscher, Benedikt M. (1991) "Effects of model selection on inference," Econometric Theory 7, 163-185.

Ptacek, Frank, and Robert M. Baskin (1996) "Revision of the CPI Housing Sample and Estimators." Monthly Labor Review (December), 31-9.

Raftery, Adrian E. (1995) "Bayesian Model Selection in Social Research." Sociological Methodology 25, 111-163.

Raftery, Adrian E., David Madigan and Jennifer A. Hoeting (1997), "Bayesian model averaging for linear regression models," Journal of the American Statistical Association 92(437), 179-191.

Sala-i-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller (2004), "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," American Economic Review, 94.4, 813-835.

Shtatland, Ernest S., Sara Moore, Inna Dashevsky, Irina Miroshnik, Emily Cain and Mary B. Barton (2000) "How to be a Bayesian in SAS: Model Selection Uncertainty in Proc Logistic and Proc Genmod." Proceedings of the 13th Annual NorthEast SAS Users Group Conference.

Stock, James H. and Mark Watson (2001) "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." In R.F. Engle and H. White (eds). Festschrift in Honour of Clive Granger, Pages 1-44.

Stock, James H., and Mark Watson (2004), "Combination forecasts of output growth in a seven country data set," Journal of Forecasting 23, 405-430.

Swanson, N.R., and T. Zeng (2001), "Choosing among competing econometric forecasts: regression-based forecast combination using model selection", Journal of Forecasting 6:425-440.

Temple, Jonathan (2000) "Growth regressions and what the textbooks don't tell you." Bulletin of Economic Research 52.3, 181-205.

Timmermann, Allan (2005) "Forecast Combinations." Mimeo, UCSD. Forthcoming, Handbook of Economic Forecasting.

van Dalen, Jan, and Ben Bode (2004) "Estimation biases in quality-adjusted hedonic price indices." Mimeo, Rotterdam School of Mgmt., Erasmus University.

Verbrugge, Randal (2007) "The puzzling divergence of rents and user costs, 1980-2004." Forthcoming Working Paper, Bureau of Labor Statistics.

Yang, Yuhong (2003) "Regression with multiple candidate models: selecting or mixing?" Statistica Sinica, 13, 783-809.

Yang, Yuhong (2004), "Combining forecasting procedures: Some theoretical results", Econometric Theory 20:176-190.

Yang, Yuhong (2005) "Consistency of Cross Validation for Comparing Regression Procedures" Mimeo, University of Minnesota.

Yuan, Zheng, and Yuhong Yang (2005) "Combining Linear Regression Models: When and How?" Journal of the American Statistical Association, Vol. 100, No. 472, December 2005 pp.1202-1214.