Implications of Time Series Models for Long-Term Projections of Health Expenditures*

April 2013

Thomas MaCurdy [1]
Peter Hansen [2]

Acumen, LLC
500 Airport Blvd., Suite 365
Burlingame, CA 94010

ABSTRACT

A variety of time series models have been proposed as frameworks for producing long-term forecasts, such as predictions of national healthcare expenditures several decades into the future which is done by government agencies and academics to assess the future financial viability of this spending. Selecting among candidate time series specifications presents an unattractive tradeoff. Stationary models are generally uninformative about long-horizon forecasts since they merely set these forecasts equal to an unconditional mean that does not distinguish recent from earlier historical shocks when constructing projections. Nonstationary time series models, on the other hand, allow for current trend deviations to affect future projection values, but they produce confidence intervals that explode as the projection horizon increases. Thus, forecasters relying on time series models face a reality wherein their predictions are either precise but uninformative, or informative but imprecise. To illustrate these challenges, the analysis estimates three time series specifications describing excess cost growth in healthcare spending and uses these estimates to compute long-term forecasts of expenditures and corresponding confidence bands. These three models span the spectrum of candidate prototypes: a stationary (short memory) model, a nonstationary integrated model, and a fractionally integrated (long memory) model. The empirical findings demonstrate that all specifications produce implausible long-term forecasts.

[1] Professor, Department of Economics, and Senior Fellow, The Hoover Institution, Stanford University, Stanford, CA 94305; and Senior Research Associate, Acumen, LLC, Burlingame, CA 94010.

[2] Professor, Department of Economics, European University Institute, Florence, Italy.

[*] This research was supported by contract RTOP CMS-08-026 from the Office of the Actuary (OACT) of the Centers for Medicare and Medicaid Services (CMS). Any opinions in this document do not necessarily represent the official position or policy of either OACT or CMS. The authors thank Todd Caldis for many useful comments.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 INTRODUCTION

Between 1985 and 2010, per capita national health expenditures (NHE) rose 6.4 percent per year; between these same years, per capita gross domestic product (GDP) rose by 4 percent per year. Whereas healthcare spending made up 10.5 percent of the economy in 1985, by 2010 this figure had reached 17.9 percent.[1] To characterize the trend of increasing health expenditures relative to GDP, academics, government agencies, and others often rely on the measure "excess cost growth," which notionally gauges the difference between the rate of healthcare spending growth and GDP growth. Researchers use the concept of excess cost growth (ECG) not only to characterize historical spending trends, but also as a convenient metric for making and presenting projections of aggregate future healthcare spending. ECG cost growth assumptions underlie long-run 75-year health spending projections produced by Centers for Medicare & Medicaid Services' (CMS) Office of the Actuary (OACT) and the Congressional Budget Office (CBO). Forecasting values of ECG into the far-distant future is an exceedingly challenging task. The purpose of this paper is to investigate the prospects of using time series model to forecast values of ECG and long-run healthcare expenditures.

A variety of time series models have been proposed in the literature as frameworks for predicting healthcare costs over the long-term horizon. The 2004 Technical Review Panel of the Medicare Trustees Reports recommended investigating and developing new approaches for modeling the long-term growth of NHE and Medicare expenditures, such as the use of time series models to create more realistic measures of forecasting uncertainty. Academic research as well (e.g., Lee and Miller (2002)) has suggested integrating time series models into existing forecasting frameworks for healthcare costs to create confidence intervals surrounding forecast values. This report catalogs the principal statistical structures of candidate time series models, presenting a succinct overview of their basic characteristics. The time series frameworks considered here are non-structural in the sense that they exploit only past realizations of the dependent variable and error terms to predict future values; they ignore the possibilities of also incorporating other explanatory variables that could assist in improving forecasts. The discussion principally focuses on depicting the core properties of time series specifications relevant in making long-run forecasts. Whereas time series frameworks have proven to be useful in forecasting annual healthcare costs for relatively short forecasting horizons, the capability of these models to formulate refined predictions far into the distance future—such as the years falling into the latter part of the 75-years Trust Fund horizon—is problematic. In particular, a forecaster faces one of two unattractive choices when specifying time series models to predict values far into the distant future: (i) Adopt a stationary specification that ignores history and merely sets all long-term forecasts equal to the same unconditional mean with fixed confidence

[1] Calculations based on National Health Expenditures Account Data without any age-gender adjustment.

band that do not increase further out in time; or (ii) Select a nonstationary specification that produces sophisticated long-horizon predictions that depend on a time-series history, but with confidence intervals that grow so large as to make predictions virtually uninformative.

To illustrate the challenges encountered in using conventional time series models to forecast healthcare costs in a long-horizon setting, this report presents estimates of three model variants that span the spectrum of candidate prototypes: a stationary (short memory) model, a nonstationary integrated model, and a fractionally integrated (long memory) model. The analysis estimates specifications measuring growth in the excess cost ratio adjusted for changing age-gender composition in the US population. The estimates from these models are then used to compute long-term forecasts of health care expenditures, along with their confidence bands. The empirical findings demonstrate that all specifications produce implausible long-term forecasts.

This report is organized into four sections. Section 2 briefly describes the forecasts of healthcare expenditures and excess cost growth produced by OACT and CBO, and provides pertinent background concerning the methods and assumptions that assist in explaining differences in their projection results. Section 3 catalogs the classes of time series models available for creating long-term forecasts of healthcare expenditures, presenting a succinct overview of their basic characteristics and depicting the core properties of these different frameworks in making long-run projections. Section 4 presents estimates for three prototype time series models describing excess cost growth in healthcare costs in the US, along with the forecasts and confidence intervals implied by these fitted models. Finally, Section 5 presents concluding comments.

2 EXCESS COST GROWTH AND LONG-RUN FORECASTS OF HEALTHCARE EXPENDITURES

To provide context for the projections of ECG produced by time series models discussed in later sections, this section describes the use of ECG in forecasting healthcare spending by OACT and CBO. Section 2.1 first explains the concept of ECG, and Section 2.2 presents the most recent 75-year Medicare projections of OACT and CBO.

2.1 Defining Excess Cost Growth

Rather than directly forecasting healthcare spending levels, many government agencies and academics make projections of health expenditures using assumptions about the long-run ECG rate. One computationally-usable formulation sets ECG equal to the difference in the rate of growth in per capita health expenditures and the rate of growth in per capita GDP, adjusted for population growth and changes in the age and gender composition of the population. This construction of ECG for year t yields:[2]

$$ECG_t = \frac{\left(\frac{HealthExp_t}{HealthExp_{t-1}}\right)}{\left(\frac{AgeGender_t}{AgeGender_{t-1}}\right)} - \left(\frac{GDP_t}{GDP_{t-1}}\right) \tag{2.1}$$

where $HealthExp_t$ represents per capita healthcare expenditures in year t; the factors $AgeGender_t$ translate per capita spending into values that occurred or would be expected to occur in year t if expenditure patterns for each age/gender grouping were the same as observed in a reference year; and GDP_t is the realized or expected per capita gross domestic product in year t. Several sources exist for calculating the $AgeGender_t$ factors incorporated in ECG projections, such as the Social Security Administration's (SSA) annual population forecasts.[3]

Drawing on lessons in the forecasting literature, one motivation for projecting ECG rather than NHE directly is that variability of ECG has been relatively more stable over time than NHE. The standard deviation of ECG between 1960 and 2010 is 2.5 percent, whereas the standard deviation of NHE growth is 3.2 percent.[4] Thus, especially if the research question primarily focusses on forecasting NHE as a share of GDP rather than NHE levels, projecting ECG can offer a more attractive option than projecting NHE.

Inspection of (2.1) reveals that ECG is a characterization of the residual growth rate of health spending after controlling for spending change driven by changes in the age-gender mix of

[2] See the appendix in Caldis (2009).
[3] See Office of the Chief Actuary of the Social Security Administration (2012) and OASDI (2012).
[4] Calculations based on National Health Expenditures Account Data without any age-sex adjustment. ECG standard deviation is also lower than the standard deviation of the GDP growth rate (2.9 percent).

the population. Also, because the ECG rate is net of the GDP growth rate, the sign of the ECG rate indicates whether the aggregate health spending evolves as an increasing or decreasing share of GDP; in particular, if ECG is positive (negative), then the share is growing (shrinking). Instead of relationship (2.1), many official references express the ECG rate as a numerically-equivalent multiplicative factor, sometimes called an xratio.[5]

2.2 OACT and CBO Long-Run Medicare Projections

Both OACT and CBO are tasked with making current-law projections of Medicare spending, which—as the name indicates—assume that all policies currently in place do not change over the entire forecast window. The goal of these projections is to provide a baseline for evaluating the financial solvency of the program as it is presently configured. A benefit of having such baselines is that, when Congress considers new policies, these agencies can project budgetary effects. In making long-term Medicare projections, both OACT and CBO have relied, and continue to rely extensively, on assumptions about future ECG rates.

OACT's excess cost growth methods have evolved over the past dozen years, but continue to be built around assumptions regarding the long-run trajectory of health spending. Long-range projections up through the Trustees' Report of 2005 were constructed using a single excess cost growth assumption of one percent above the projected rate of GDP growth (GDP+1), assumed to prevail for all parts of the U.S. health sector.[6] Starting with the Trustees Report of 2006 and continuing through the 2011 report, the GDP+1 assumption was implemented in conjunction with a computable general equilibrium (CGE) model applied to produce a decelerating series of projected annual excess cost growth rates, with the present value of the 75-year actuarial balance made equivalent to the simple GDP+1 projection.[7] A decelerating series of ECG was regarded as a more plausible characterization of transition to a long-run steady state of roughly GDP+0, the ultimate long-range assumption of the Medicare Trustees. In response to recommendations of the 2010-2011 Medicare Technical Panel, in 2012 the Trustees dispensed with the CGE model and started using a Factors Contributing to Growth Model in conjunction with ECG assumptions to produce the long-range projections appearing in the 2012 Report of the Medicare Trustees.[8]

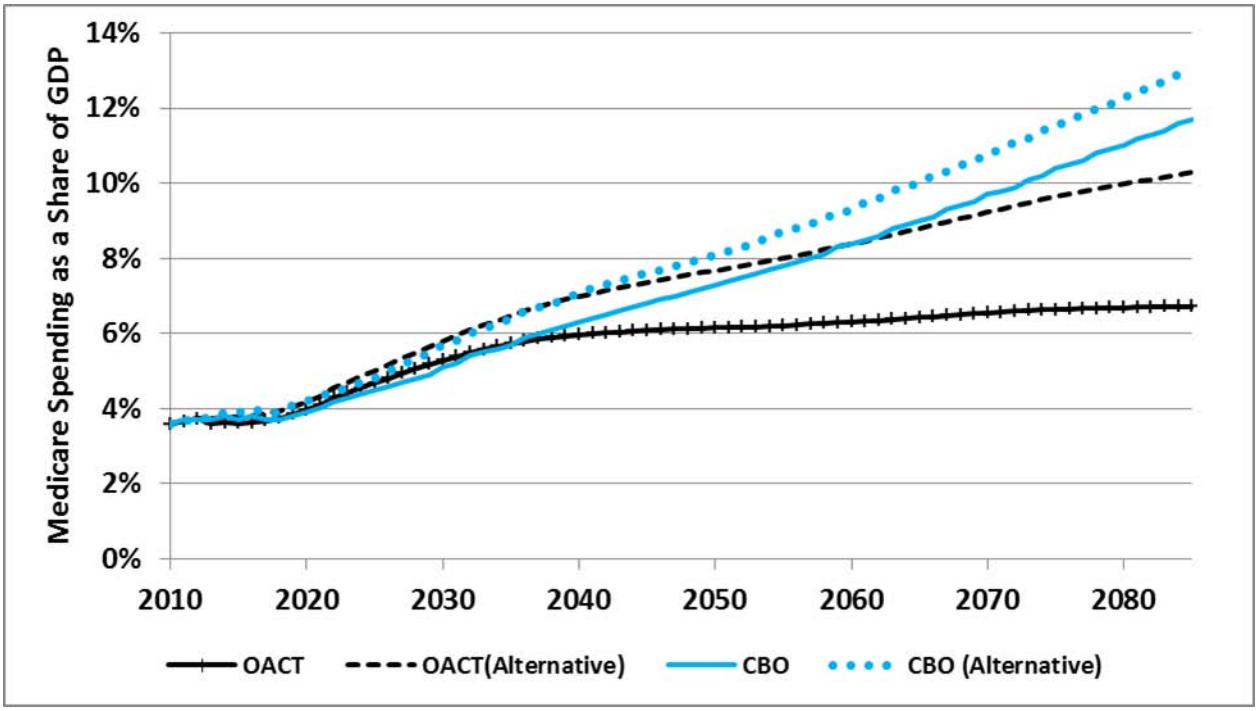
CBO has traditionally implemented its long-range projections by adopting different cost growth rate assumptions for the three segments of the U.S. health sector: Medicare, Medicaid, and the rest of the healthcare market. Historical time series data are used to compute an initial long-range cost growth rate, and rates of deceleration in cost growth are then assumed for each

[5] Caldis, (2009)
[6] Caldis (2009)
[7] Caldis (2009)
[8] Heffler, Caldis, Smith (2012)

segment of the U.S. health sector. In its 2012 long-range projections CBO assumed that Medicare will continue to grow more rapidly than Medicaid and the rest of the U.S. health sector in the long-run, with ECG in Medicare only decelerating to GDP+1 by the 75th projection year. This higher rate of Medicare cost growth maintained by CBO leads to a divergence in the current-law projections reported by OACT and CBO for Medicare in the very long run.

Figure 2.1 displays OACT's and CBO's most recent long-run Medicare spending projections.[9] Although OACT's current-law Medicare projection is similar to CBO's over the medium-term, its long-run Medicare spending projections fall well below those of CBO. In 2035, OACT's current projection estimates Medicare spending will make up 5.6 percent of the economy; CBO's estimate in that same year is 5.7 percent. By 2080, however, the projections diverge sharply. OACT projects that Medicare spending will comprise 6.7 percent of GDP, whereas CBO's baseline projection estimates that Medicare expenses will be 11.0 percent of GDP.

Figure 2.1: OACT and CBO Projections of Medicare Spending for 2010-2085



In addition to these current-law projections, both organizations also build alternative projections relying on more realistic assumptions. OACT researchers acknowledge that their

[9] Data from the figure come from two sources: (i) CBO LBTO supplemental data 2012; and (ii) and OACT memoranda from Shatto and Clemens 2012. The OACT data are smoothed across years. ECG assumptions used by OACT and CBO apply only to long-range time horizons. For example, OACT produces 10-year bottom-up short-range projections, a set of intermediate horizon projections for years 11 through 24 that transition from short-range to long-range rates, and long-range methods are implemented for years 25 through 75 of the long-range projection horizon. CBO projections also distinguish between short-range and long-range methods.

current-law projections are "clearly unrealistic" with respect to certain assumptions, especially physician expenditures.[10] OACT's alternative model differs from its extended baseline scenario along three dimensions. The alternative model assumes that (i) scheduled decreases to physician payments (i.e., Sustainable Growth Rate (SGR) reductions) do not occur, (ii) reductions to various prospective payment system updates based on economy-wide productivity are assumed to phase out and end by 2034, and (iii) the Independent Payment Advisory Board (IPAB) requirements are not implemented.[11] CBO adopts a similar set of assumptions for their extended alternative fiscal projection. In particular, the CBO long-run alternative projection assumes that "ongoing reductions in payment updates for most providers in the fee-for-service program, the Sustainable Growth Rate mechanism for payment rates for physicians, and the IPAB..." will not continue into the future.[12] Under these alternative sets of assumptions, OACT projects that Medicare spending will be 10.0 percent of the economy by 2080 and CBO projects that Medicare spending will make up 12.3 percent of GDP in that same year. Figure 2.1 displays these alternative projections along with the current-law forecasts.

[10] See Shatto and Clemens 2012, p.1.
[11] See Shatto and Clemens 2012.
[12] See CBO LTBO 2012

3 PROPERTIES OF FORECASTS USING TIME SERIES MODELS

Using time series models to forecast ECG offers a natural alternative to the deterministic approaches currently used by OACT and CBO. The following discussion catalogs the principal statistical structures of these time series models, presenting a succinct overview of their basic characteristics. The discussion primarily focuses on depicting the core properties of these different frameworks in making long- run forecasts. All prominent time series models possess a familiar moving average (MA) representation. Indeed, translating the different time series structures into this common representation offers a rich framework for discussing and comparing the properties of forecasts across models. This is the approach adopted throughout this report.

The following discussion describes the framework underlying the forecasts produced by time series models. Section 3.1 presents the core MA representation of time series models, and Section 3.2 summarizes the measures of uncertainty implied for forecasts produced by alternative structures of this MA process. Section 3.3 presents the MA structures for all the primary specifications of time series processes found in the literature. Section 3.4 identifies the unattractive tradeoffs inherent in selecting time series specifications linked to long-horizon forecasting, and, finally, Section 3.4 illustrates the numerical relevance of these tradeoffs.

3.1 Predicting Future Values Based on MA Models

To fix ideas, let $\hat{X}_{T+h,T}$ denote a forecast of X_{T+h} that is made at time *T*. The optimal forecast depends on the loss function. With the mean square error loss function, the most popular choice, the optimal forecast is the conditional expectation, $E_T(X_{T+h})$, and the expected prediction loss is directly given by the prediction error variance. With a more complex loss function, the optimal forecast will depend on additional features of the distribution of $X_{T+h} - E_T(X_{T+h})$. Given the widespread popularity of quadratic loss functions, the following discussion focuses on identifying how various time series model decompose X_{T+h} into the conditional mean, $E_T(X_{T+h})$ and the difference $X_{T+h} - E_T(X_{T+h})$. For simplicity, we refer to the former as the forecast and the latter as the prediction error, even though this is only optimal under mean square error loss.

To develop multi-period-ahead forecasts of X_t assuming it follows a time series model, a wide variety of time series can be expressed as an infinite order moving-average process (i.e., $MA(\infty)$):

$$X_t = \mu_t + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} \equiv \mu_t + \upsilon_t \tag{3.1}$$

where μ_t is a deterministic term, $\varepsilon_t = X_t - E(X_t \mid X_{t-1}, X_{t-2}, ...) = X_t - E_{t-1}(X_t)$ with $\{\varepsilon_t\}$ being a white noise process (i.e., the $\varepsilon_t's$ are serially uncorrelated errors), and the random variable υ_t measures the accumulative influence of idiosyncratic errors. The Wold representation theorem states that any covariance stationary process can be written as a *MA*(∞) process in the form of (3.1),[13] and Beveridge-Nelson (1981) show that integrated processes of order one (i.e., models with unit roots) have analogous MA representations.[14] Using MA representation (3.1), the expected value of X_{T+h} given the history $X_T, X_{T-1}, ...$ is:

$$\hat{X}_{T+h,T} = E_T(X_{T+h}) = E(X_{T+h} \mid X_T, X_{T-1}, ...) = E(X_{T+h}) + \sum_{j=h}^{\infty} \psi_j \varepsilon_{T+h-j}. \tag{3.2}$$

This is the natural *h*-period ahead forecast, as it minimizes the variance of the prediction error

$$\sum_{j=0}^{h-1} \psi_j \varepsilon_{T+h-j} \tag{3.3}$$

3.2 Measuring Uncertainty of Forecasts

The principal difference in the various formulations of time series models concerns whether the coefficients ψ_k in (3.1) converge to zero as *k* increases (i.e., $\psi_k \to 0$ as $k \to \infty$) and the rate at which such convergence occurs. If $\psi_k \to 0$ as $h \to \infty$ sufficiently fast, then the process is stationary. If not, the time series process is nonstationary. Two prototype specifications of a time series process illustrate the factors governing the rate of convergence of the ψ_k coefficients.

One prototype takes the form

$$(1 - \varphi L)\, \upsilon_t \;=\; \varepsilon_t \;. \tag{3.4}$$

Inverting $(1-\varphi L)$ yields

$$(1-\varphi L)^{-1} = 1 + \varphi L + \varphi^2 L^2 + \cdots = \sum \psi_j L^j,$$

which permits expressing (3.4) as

$$\upsilon_t = (1-\varphi L)\varepsilon_t = \sum_{j=0}^{\infty} \psi_{h+j} \varepsilon_{t-j} \;. \tag{3.5}$$

This relationship corresponds to a *MA*(∞) process with coefficients ψ_k decaying (growing) at an exponential rate φ^j. When $\varphi < 1$, these coefficients decay and the process is stationary, and when $\varphi \geq 1$, the process is nonstationary. More generally, this prototype captures the essential features of the familiar finite-order autoregressive moving-average (ARMA) model commonly

[13] Abstracting from technicalities having to do with ergodicity.

[14] The Beveridge-Nelson representation can be written in the form: $X_t = \mu_t + X_0 + \sum_{j=0}^{t-1} \psi_j \varepsilon_{t-j}$

used in the literature. The quantity $(1-\varphi L)$ in (3.5) corresponds to the autoregressive (AR) component of the model. An ARMA specification falls into the short-term memory class and is stationary when all the roots of its AR component are less than one in absolute value; the ARMA model falls into the integrated class and is nonstationary when any of the roots of its autoregressive component equal one in value and when the others are less than or equal to one; and the model is also nonstationary when any of its autoregressive roots exceed one.

A second prominent prototype specification of a time series process takes the form

$$(1-L)^d \upsilon_t = \varepsilon_t \tag{3.6}$$

where L is the "lag" operator (i.e., $LX_t = X_{t-1}$), and d represents the "order of integration." Inverting $(1-L)^d$ yields [15]

$$(1-L)^{-d} = 1 + d\cdot L + \frac{1}{2}d(d+1)L^2 + \cdots = \sum \psi_j L^j,$$

which implies a rewriting of (3.6) given by

$$\upsilon_t = (1-L)^{-d}\varepsilon_t = \sum_{j=0}^{\infty} \psi_{h+j}\varepsilon_{t-j} \ . \tag{3.7}$$

This relationship corresponds to a $MA(\infty)$ process with coefficients ψ_k decaying (growing) at a geometric rate; in particular, for large integers j,

$$\psi_j \cong (j+1)^{d-1} \ .$$

For $d = 0$, relationship (3.7) represents a "short-memory" process; for $0 < d < 1$; this relationship corresponds to what is known in the literature as a "long-memory" (or "fractionally integrated") process; and for $d = 1$, relationship (3.7) becomes an integrated process (i.e., a process with unit roots). Time series (3.6) is stationary when $d < \frac{1}{2}$; and it is nonstationary when $d \geq \frac{1}{2}$.

To characterize properties of the uncertainty in forecasting a h-period ahead value based on the $MA(\infty)$ process (3.1), the future value of υ_t measured h periods ahead equals:

$$\begin{aligned} \upsilon_{t+h} &= \sum_{j=0}^{\infty} \psi_j \varepsilon_{t+h-j} \\ &= \sum_{j=0}^{h-1} \psi_j \varepsilon_{t+h-j} + \sum_{j=h}^{\infty} \psi_j \varepsilon_{t+h-j} \ . \end{aligned} \tag{3.8}$$

[15] The inverse of $(1-L)^d$ is obtained from the Taylor expansion of $f(z) = (1-z)^{-d}$ which yields $f(z) = 1 + d\cdot z - \frac{1}{2}d(d+1)z^2 + \cdots$.

The minimum-variance unbiased prediction of v_{t+h} conditioning on information available in period *t* equals:

$$E_t(v_{t+h}) = \sum_{j=h}^{\infty} \psi_j \varepsilon_{t+h-j} \quad . \tag{3.9}$$

Substituting (3.9) into (3.8) produces

$$v_{t+h} = \underbrace{E_t(v_{t+h})}_{\text{Forecast}} + \underbrace{\sum_{j=0}^{h-1} \psi_j \varepsilon_{t+h-j}}_{\text{Forecast Error}} \quad . \tag{3.10}$$

Consequently, the variance of the forecast error becomes:

$$\sum_{j=0}^{h-1} \psi_j^2 \sigma_\varepsilon^2 \quad . \tag{3.11}$$

Clearly one sees that the properties of this variance of the forecast depend directly on features of the coefficients ψ_k as *k* increases. For a stationary process, this variance converges to a constant as *h* grows large (i.e., as $h \to \infty$), which requires

$$\sum_{j=1}^{\infty} \psi_j^2 < \infty \quad . \tag{3.12}$$

For a nonstationary process, this variance grows steadily with the length of the forecast horizon *h* with the rate of increase related to how large the coefficients ψ_k become as *k* increases. The variance expression in (3.12) is used in the construction of statistically-derived confidence intervals for forecasts.

3.3 Classifications of Familiar Time Series Processes as MA Models

All major specifications of time-series processes throughout the literature have moving average representations with structures characterized by one of the two prototype formulations described above. The following discussion categorizes the most popular specifications and shows their MA forms to exhibit their forecasting properties. The analysis begins with stationary short- and long-memory models, and then turns to several nonstationary variants.

3.3.1 Stationary Short-Memory Processes

The workhorse of times series forecasting is the autoregressive moving average model, a framework known as the Box-Jenkins methodology. An *ARMA*(*p*,*q*) model takes the form

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \rho + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \tag{3.13}$$

where the univariate time series, X_t, depends linearly on its own recent past as well as the innovation (shock) ε_t, and its recent past. An ARMA process has a stationary representation if all solutions to

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

are greater than one in absolute value.

Any stationary ARMA process can be expressed as an infinite order moving average process of the form (3.1), where its MA coefficients can be derived recursively from the Yule-Walker equations:

$$\begin{aligned}
\psi_0 &= 1 \\
\psi_1 &= \phi_1\psi_0 - \theta_1 \\
\psi_2 &= \phi_1\psi_1 + \phi_2\psi_0 - \theta_2 \\
&\vdots \\
\psi_k &= \phi_1\psi_{k-1} + \cdots + \phi_p\psi_{k-p} - \theta_k
\end{aligned}$$

with $\theta_k = 0$ for $k > q$. Solving these equations reveals that ψ_k is of the same order as $|\lambda|^k$ where λ is the solution to $\phi(z) = 0$ with the largest absolute value; consequently, $\psi_k \approx |\lambda|^k$ for some $|\lambda| < 1$. This exponential rate of decay is quite rapid.

For this specification of the MA process, the long-term forecast $\hat{X}_{T+h,T}$ quickly converges to $E(X_{T+h})$ as the prediction horizon, *h*, reaches even modest values, and the variance of the associated prediction error approximately equals $\sum_{j=0}^{\infty} \psi_j^2 \sigma_\varepsilon^2$. Since these quantities represent the unconditional mean and variance of the X_t time series process, the particular specification of the ARMA model and its parameters describing dynamics have little influence on either the long-term forecast or its variance.

3.3.2 Adding a GARCH Structure

A GARCH model introduces time variation in the conditional variance $var_T(\varepsilon_{T+1}) = \sigma_{T+1}^2$. For instance, a GARCH(1,1) model assumes

$$\sigma_t^2 = \gamma + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \,. \tag{3.14}$$

This in turn implies the following *AR(1)* structure for the conditional variance,

$$\sigma_t^2 = \gamma + (\alpha + \beta)\sigma_{t-1}^2 + \xi_t \qquad \text{with} \qquad \xi_t = \alpha(\varepsilon_{t-1}^2 - \sigma_{t-1}^2) \,.$$

Also, $E_t(\xi_{t+1}) = 0$.

The primary purpose of proposing a GARCH framework is to capture volatility dynamics in the short run. In financial time series, empirical estimates often suggest $\alpha + \beta \approx 1$, which indicates highly persistent volatility—resembling a near-unit root process.

However, a GARCH model does not affect the conditional mean $E_T(X_{T+h})$. Moreover, volatility dynamics average out in the long-run—assuming $|\alpha + \beta| < 1$ so that $E(\sigma_t^2) = \frac{\gamma}{1-\alpha-\beta}$. So, the long-term prediction error variance, $var_T(X_{T+h})$, is not affected either.

3.3.3 *Stationary Long-Memory Processes*

A prominent category of long-memory models is known as fractionally integrated processes. A simple representation of such process takes the form of model (3.6) with the difference operator d in the range $0 < d < 1$. A fractionally integrated process with $0 < d < \frac{1}{2}$ is stationary with coefficients, ψ_i, converging to zero at a geometric rate that is slower than exponential decay associated with stationary short-memory processes. Consequently, fractionally integrated processes are more dependent on ε's in the distant past, than is the case for ARMA processes. This motivates their designation as *long memory processes.*

Using the above results describing the second prototype time series model, the forecast of a fractionally integrated process can be expressed in a MA representation given by

$$\hat{X}_{T+h,T} = \sum_{j=0}^{\infty} \psi_{h+j} \varepsilon_{T-j}$$

where the coefficients $\psi_j \cong (j+1)^{d-1}$ for large integers j. The key difference between this formulation and the one obtained for the autoregressive processes is that the prediction error variance will tend to be larger for the long-memory models. In particular,

$$var(X_{T+h} - \hat{X}_{T+h,T}) = \sigma_\varepsilon^2 \sum_{j=0}^{h-1} \psi_j^2 \cong \sigma_\varepsilon^2 \sum_{j=0}^{h-1} (j+1)^{2(d-1)}$$

This variance remains finite as $h \to \infty$ when $d < \frac{1}{2}$. While it takes longer for this expression to attain the value of the unconditional variance of X_t than occurs for short-memory processes, it is still true that this convergence takes place fairly rapidly. The implication is again that the predictive distribution for a h-periods ahead forecasts reduces to the stationary distribution when h is large.

3.3.4 *Nonstationary Long-Memory Processes*

A fractionally integrated process with the difference operator d in the range $\frac{1}{2} \le d < 1$ is nonstationary. In this instance, the coefficients, ψ_i, in the MA representation of this process either do not converge to zero or converge at a sufficiently slow rate so that the variance grows continually as h increases and explodes as $h \to \infty$. In fact, the width of a confidence interval for a forecast will expand at the rate $h^{d-\frac{1}{2}}$, where h is the forecasting horizon. This follows from the known result

$$h^{\frac{1}{2}-d} \sum_{j=1}^{h} [\Delta X_{T+j} - E(\Delta X_{T+j})]d \to C \int_0^1 (1-u)^{d-1} dW(s)$$

where $W(s)$ is a Brownian motion, and C is a constant. The integral on the right hand side is known as a functional Brownian motion of type II.

3.3.5 *Integrated Processes*

Integrated models easily constitute the most popular statistical apparatus for describing processes that exhibit persistence in the long run. These models encompass autoregressive structures with unit roots. Referring to the autoregressive function $\phi(z)$ given by (3.2), integrated structures are those for which the solution to the equation $\phi(1) = 0$ has one or more unit root. One finds several variants of unit root processes, each characterized by their order of integration.

By far, the most-studied and relevant type of unit root processes are those integrated of order one, designated as $X_t \sim I(1)$. The $I(1)$ processes are characterized by the requirement that the difference, $\Delta X_t = X_t - X_{t-1}$, has a $MA(\infty)$ structure with exponentially decaying weights and $\sum_{j=0}^{\infty} \psi_j \neq 0$. One can express such processes as

$$X_t = c\sum_{s=1}^{t} \varepsilon_s + c(L)\varepsilon_t + Y_0 \tag{3.15}$$

where $Y_0 = f(X_0, X_{-1}, ...)$ is an initial value and $c(L)\varepsilon_t = c_0\varepsilon_t + c_1\varepsilon_{t-1} + \cdots$ is a time-invariant $MA(\infty)$ process. Expression (3.15) constitutes the Beveridge-Nelson (BN) representation of the process. It is the initial value, Y_0, and random walk component, $\sum_{s=1}^{t} \varepsilon_s$, that make this representation distinctly different from the stationary $MA(\infty)$ representation obtained previously.

The conditional mean X of X_{T+h} is

$$\hat{X}_{T+h,T} = \sum_{j=h}^{\infty} c_j \varepsilon_{T+h-j} + Y_0; \tag{3.16}$$

and the prediction error is

$$\begin{aligned} X_{T+h} - \hat{X}_{T+h,T} &= c\sum_{j=1}^{h} \varepsilon_{T+j} + \sum_{j=0}^{h-1} c_j \varepsilon_{T+h-j} \\ &= \sum_{j=0}^{h-1} \left(c + c_j \right) \varepsilon_{T+h-j} \end{aligned} .$$

Since $c_j \to \infty$ at exponential rate, the random walk term, $c\sum_{j-1}^{h} \varepsilon_{T+j}$, will be the dominating factor in long-horizon forecasts. For large h, the prediction error variance becomes

$$var(X_{T+h} - \hat{X}_{T+h,T}) \approx hc^2\sigma_\varepsilon^2.$$

So, unit root processes have the unattractive feature that the prediction error variance increases linearly with the horizon.

The following example presents a unit root process and provides its BN representation. Consider the $AR(2)$ process

$$X_t = \frac{1}{2} X_{t-1} + \frac{1}{2} X_{t-2} + \varepsilon_t.$$

Since $\Delta X_t = -\frac{1}{2}\Delta X_{t-1} + \varepsilon_t$, we immediately see that the $MA(\infty)$ representation for ΔX_t takes the form:

$$\Delta X_t = \sum_{i=0}^{\infty} \left(-\tfrac{1}{2}\right)^i \varepsilon_{t-i} .$$

Further, it can be shown that

$$X_{T+h} = \frac{2}{3} \sum_{t=T+1}^{T+h} \varepsilon_t + C(L)\varepsilon_{T+h} + \frac{1}{3}[2X_T - X_{T-1}],$$

with $c_0 = \frac{1}{3}$, $c_1 = -\frac{1}{6}$, $c_2 = \frac{1}{12}$, and $c_k = \frac{1}{2}(c_{k-1} + c_{k-2})$ for $k \geq 2$. Thus,

$$X_{T+h} - E_T(X_{T+h}) = \frac{2}{3} \sum_{t=T+1}^{T+h} \varepsilon_t + \left(\frac{1}{3}\varepsilon_{T+h} - \frac{1}{6}\varepsilon_{T+h-1} + \cdots + c_{h-1}\varepsilon_{T+1} \right).$$

Computing the variance of this expression reveals $var(\sum_{t=T+1}^{T+h} \varepsilon_t) = h \cdot \sigma_\varepsilon^2$, and $var(\frac{1}{3}\varepsilon_{T+h} - \frac{1}{6}\varepsilon_{T+h-1} + \cdots + c_{h-1}\varepsilon_{T+1}) \to K$(a constant) as $h \to \infty$. It is clear that the random walk term (often called the stochastic trend), $\frac{2}{3}\sum_{t=T+1}^{T+h} \varepsilon_t$, constitutes the dominating factor in the prediction errors.

3.3.6 *Exponential Smoothing*

Finally, a simple and popular method for forecasting is exponential smoothing where the prediction of X_{T+1} at times T is an exponentially weighted average of past observations

$$\hat{X}_{T+1,T} = (1-\rho)\sum_{j=0}^{\infty} \rho^j X_{T-j}$$

for some $0 \leq \rho < 1$. The distant past is heavily discounting when ρ is close to zero, and influential when ρ is close to one. An elegant feature of exponential smoothing is that its predictions are simply updated by the recursion

$$\hat{X}_{T+1,T} = \rho \hat{X}_{T,T-1} + (1-\rho) X_T.$$

This exponential smoothing structure implicitly assumes that X_t follows a $ARMA(1,1)$ process with a unit root in the autoregressive component process. More specifically,

$$X_t = X_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}.$$

This is also called a $IMA(1)$ model.

Another label for this process is "local level model." This model carries the interpretation that X_t is a noisy observation of a latent process, Y_t, where Y_t is a simple random walk. The formulation capturing these relationships is given by:

$$X_t = Y_t + \eta_t$$
$$Y_t = Y_{t-1} + \upsilon_t$$

where η_t and υ_t are both iid and mutually independent.

3.4 Unattractive Tradeoffs in the Selection of Time Series Models

Inspection of formulas (3.2), (3.11) and (3.12) reveals an inherent tradeoff in the properties of the long-run forecasts produced by MA process (3.1). According to (3.2), the extent to which history of the time series at time *T* matters in the long-term predictions $\hat{X}_{T+h,T}$ depends on the magnitudes of the ψ_k's for values of *k* above *h*. If the ψ_k's decay rapidly, then shocks in X_t prior to *T* dissipate quickly and are highly discounted in $\hat{X}_{T+h,T}$. In contrast, if the ψ_k's persist and either converge to zero very slowly or not at all, then shocks in X_t prior to *T* can influence values of the forecasts $\hat{X}_{T+h,T}$ far into the future.

At the same time, expression (3.11) for the variance of forecasts shows that the size of confidence intervals for future predictions also directly depends on the convergence rate of the ψ_k's. If these coefficients converge quickly to zero, then confidence intervals for forecasts are small and bounded. Alternatively, if the ψ_k's decay slowly or not at all, then the prediction variance is large and may not even be bounded.

Consequently, a forecaster faces the following unattractive choices when using time series models to predict values far into the distant future. A forecaster could adopt a specification featuring a fast decay of the ψ_k's. For such a model, the unconditional distribution holds all the available information about X_t in the distant future. The implied long-term prediction is the unconditional mean of the process, and the confidence interval is fully determined by the unconditional variance of the process. Knowledge about the dynamics of X_t is only useful to the extent that it can help characterize the unconditional distribution, when the objective is to make long horizon predictions (i.e., when *h* is large).

Alternatively, a forecaster could take up a specification featuring either a slow decay or nonconvergence of the ψ_k's. Such specifications always arise for nonstationary processes. The resulting models produce sophisticated long-horizon predictions in the sense that these forecasts depend on the history of the time series process along with the dynamic properties of the model. The forecasts are not merely the unconditional mean of the distribution. However, the confidence intervals for these long-term projections can be quite large and steadily grow the

longer the horizon. The implication for these models is that little can be said about the distant future, because the uncertainty associated with any forecast will be large when *h* is large.

This tradeoff leaves unappealing options for using standard time series models to formulate forecasts and corresponding confidence intervals when faced with the problem of predicting healthcare costs several or many decades ahead. Consider the forecasting problem of predicting *h* periods ahead in the situation where *h* is large. If a researcher relies on a stationary model, then long-run forecasts merely converge to an unconditional distribution with a constant variance (regardless of the length of the horizon) and no influence of current values on forecasts. This is true irrespective of whether one considers short or long memory models. In such instances, the particular specification of the time series for the idiosyncratic errors is essentially irrelevant for long-term forecasts. Alternatively, if a researcher adopts a nonstationary model, the properties of its long-run forecasts depend on its order of integration, designated by *d* in the first prototype model introduced above. Higher *d* means that current and recent history figure more into long-run forecasts, but with the confidence intervals for predictions that grow at the rate $h^{d-\frac{1}{2}}$. Thus, the more that current values in a nonstationary models say about forecasts in the long-term future, the more uncertainty that the model associates with these forecast and this uncertainty become quite large the further out one projects outcomes.

At first impression, one might surmise that these challenges can be avoided by predicting the average growth rate of a time series rather than its future levels. However, the statistical properties associated with predicting average growth fully carry over to the problem of forecasting levels. These two approaches are in fact equivalent since one can readily convert one forecast into the other, and the same is true for the confidence intervals associated with the two methods of prediction. Whereas the confidence intervals for the average growth rate tend to shrink as the horizon increases—e.g., in the case of unit root models these intervals converge to a fixed width—the stabilization of confidence intervals for the averages portrays a false sense of certainty about the distant future projections for levels. Regardless of how they are constructed, forecasts of levels have statistical properties as outlined in the earlier discussion.

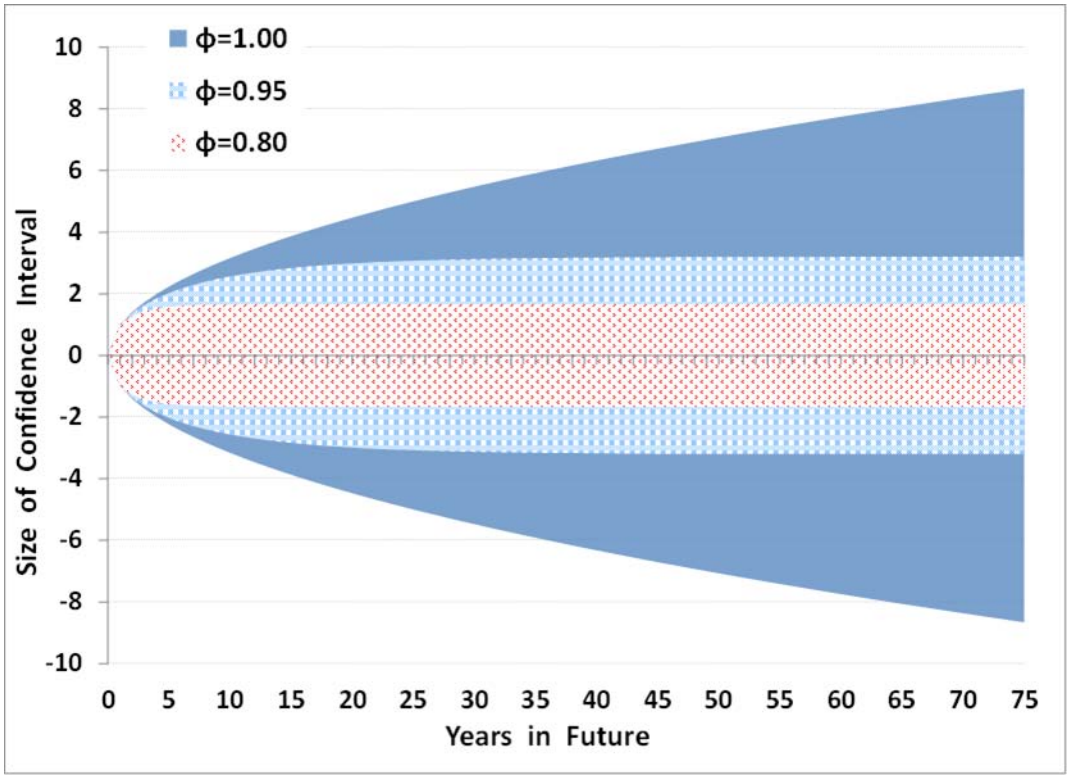
In time-series modeling as in other approaches of forecasting, prediction errors arise from two distinct sources: the first is the intrinsic source described above linked to the structure of the time series model, and the second is the uncertainty about the true data-generating process related to estimation error. A forecaster attempts to mitigate this second form of uncertainty through the choice of estimation method and the use of as much data as possible. A forecaster's goal is to minimize the influence of this factor, namely, finding a good prediction model and obtaining accurate estimates of its unknown parameters. The researcher cannot do anything about the first source of uncertainty for it is an inherent property of the time series model. In practice there is uncertainty about the choice for model, and unknown parameters must be

estimated. In practice, then, the prediction accuracy will typically be worse than the one that is found in the hypothetical situation where the true data generating process is known.

3.5 Illustrating the Tradeoffs Encountered in Long-Term Forecasts

To illustrate the challenges encountered in using time series models to project long-run healthcare costs, Figures 3.1, 3.2 and 3.3 present findings for three illustrations of the two prototype models discussed above. Turning initially to the first prototype (3.4), one sees that this is simply an *AR(1)* model. Figure 3.1 illustrates the width of the confidence interval for a forecast as a function of the prediction horizon.[16] The figure presents intervals for coefficient values $\varphi = 1.00$, $\varphi = 0.95$ and $\varphi = 0.80$. When $|\varphi| < 1$ the process is stationary, whereas $\varphi = 1$ defines a random walk which is a nonstationary unit-root process.

Figure 3.1: Size of Confidence Intervals for AR-Type Models

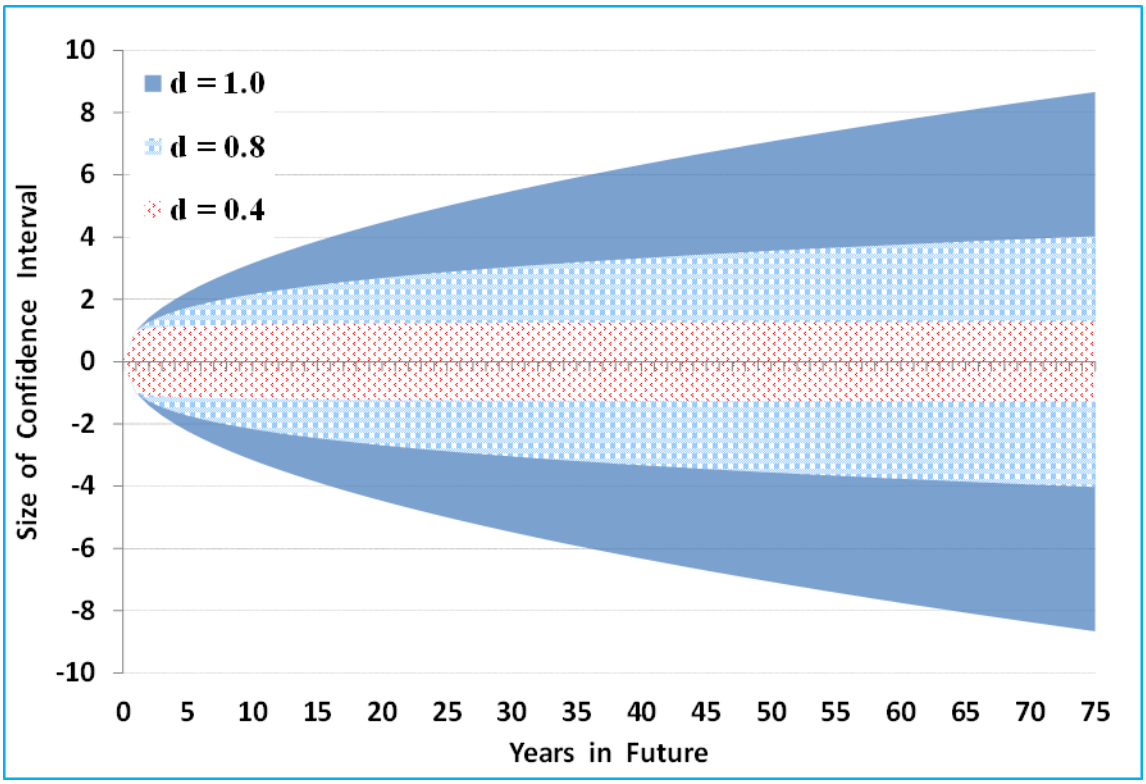


Analogous to Figure 3.1, Figure 3.2 shows how much a prediction confidence interval grows by the length of the forecast horizon *h* for the second prototype (3.6) incorporating the

[16] Whereas the width of these intervals are not too meaningful, their relative width can be compared across specifications. With C_α denoting the critical value, the width of the confidence interval for an *AR(1)* process equals $2 \bullet \sigma_\varepsilon \bullet C_\alpha \bullet \sqrt{1+\varphi^2+\varphi^4+\cdots+\varphi^{2(h-1)}}$. The computations in Figure 3.1 set, $2 \bullet \sigma_\varepsilon \bullet C_\alpha = 1$ an arbitrary normalization.

long-memory time series specifications. The process is stationary for the case when *d=0.4*, and the prediction confidence interval converges to the size defined by the unconditional variance. For the nonstationary cases *d=0.8* and *d=1*, the prediction error variance grows without bound, and the width of the prediction confidence interval grows at the rate $h^{d-\frac{1}{2}}$.

Figure 3.2: Size of Confidence Intervals for Long-Memory Models



For the stationary models, the width of the confidence intervals quickly levels off as the horizon increases. While at first impression this might seem appealing, this convergence of the confidence interval merely reflects that the best prediction at longer horizons is essentially given by the unconditional stationary distribution of v_t. This in turn means that current information has no value for making long-horizon forecasts when the model is stationary. For the nonstationary specification, the confidence interval grows without bound and rapidly so. For a random walk specification, the current value of the process is fully informative about all future forecasts since these forecasts are simply the current value. However, the implied range of uncertainty at long horizons becomes so wide as to make the projection virtually uninformative.

A convenient way to quantify the relationship between the effects of the current values of a time series on future forecasts and the uncertainty linked to these forecasts is to measure the *Prediction Signal-to-Noise ratio* (PSN). For the moving average process (3.1) and its variance (3.11), a one-standard deviation "shock" today will induce an adjustment in the prediction of v_t

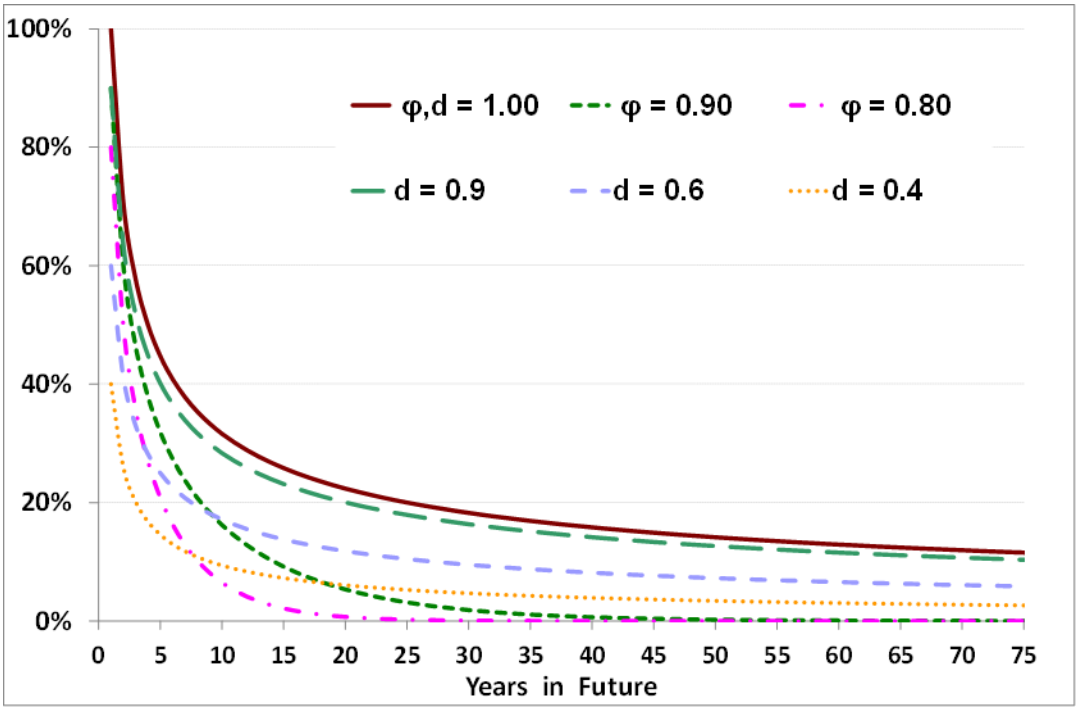
by $\sigma_\varepsilon \times \psi_h$, and the width of the corresponding prediction confidence interval is given by $\sqrt{\sum_{j=0}^{h-1} \psi_j^2 \sigma_\varepsilon^2}$. This motivates the PSN defined by

$$PSN_h = \frac{\psi_h}{\sqrt{\sum_{j=0}^{h-1} \psi_j^2}}, \qquad h = 0,1, \ldots,$$

Formally, PSN summarizes the effect of today's news v_t on the *h*-period ahead prediction relative to the uncertainty associated with this prediction.

Figure 3.3 displays the PSN for $\varphi = 1$ and $\varphi = 0.80$ specifications of the *AR(1)* processes shown in Figure 3.1, and also includes the specification $\varphi = 0.90$. In addition, this figure presents the PSNs for $d = 0.40$ $d = 0.60$ and $d = 0.90$ specifications of the long-memory processes shown in Figure 3.2. Whereas integration order $d = 0.40$ specifies a stationary long memory process, integration orders $d = 0.60$ and $d = 0.90$ designate two non-stationary long memory processes. Note that the figure also includes specification $d = 1$ (i.e., a nonstationary unit-root process) since the *AR(1)* model $\varphi = 1$ defines this case.

Figure 3.3: Signal-to-Noise Ratio as a Function of Prediction Horizon



The decay patterns of the PSN reinforce the implications of the above depictions of confidence intervals. In stationary models the signal quickly diminishes, while the noise stays

constant. The opposite situation arises for the unit root case; the signal stays constant and the noise increases linearly with the horizon. Although nonstationary models are attractive in that their current values and recent history influence future projections, the confidence intervals produced by these models grow at a sufficiently rapid rate so as to render any current data uninformative about long-term projections.

The above findings reveal an inability of standard time-series models to use current information in a series to improve long-horizon forecasts without simultaneously sharply increasing the uncertainty associated with these predictions. One encounters a *Catch 22* property. A time-series model specifies the dynamic properties of a data-generating process, and these dynamic properties dictate how news about the process today will impact the process in the future. For a long-horizon forecast to be informed by current data, it is necessary that today's news matter for the future. However, if information available today is important for the distant future, then future information (that becomes available after the forecast is made) must also be important for the distant future. Since the parameters determining these informational impacts are the same as those determining the confidence intervals of these forecasts, higher informational impact on a long-term prediction necessarily means greater uncertainty. Moreover, the structure of conventional time series models implies that uncertainty grows at a rapid rate relative to informational content, and this uncertainty continues to compound relentlessly the further away the prediction. This property makes time series models unattractive candidates for long-horizon forecasting of the sort done by OACT and CBO in their projections of healthcare expenditure in 25, 50 or 75 years from now.

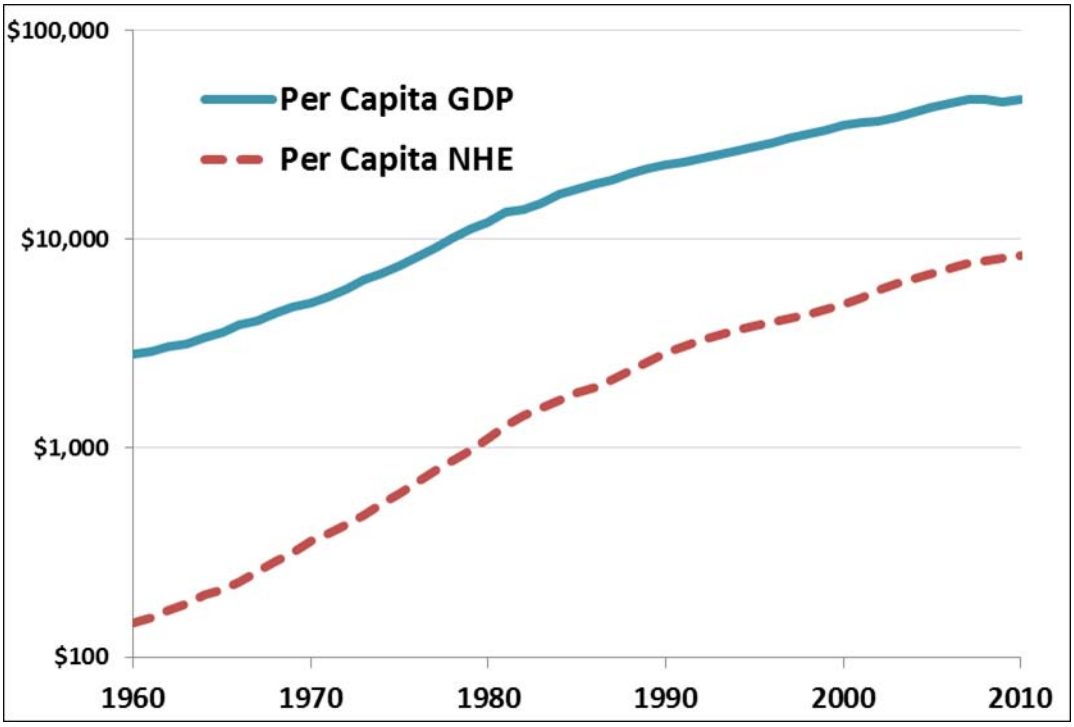
4 EMPIRICAL APPLICATION: FORECASTS OF EXCESS COST GROWTH

The following discussion illustrates the empirical properties of long-term forecasts of US expenditure on healthcare produced by the three principal variants of time series models described above: a stationary (short memory) model, a nonstationary integrated model, and a fractionally integrated (long memory) model. The analysis estimates specifications using data on a measure of the excess cost ratio adjusted for changing age-gender composition in the US population. The estimates from these models are used to compute long-term forecasts for healthcare expenditure, along with their confidence bands.

4.1 Description of Data

The data used in this empirical analysis on NHE and GDP are the same time series used by OACT and CBO in their long-term projections of healthcare expenditures. Figure 4.1 plots the per capita values of these series over the 1960 to 2010 period in the US.

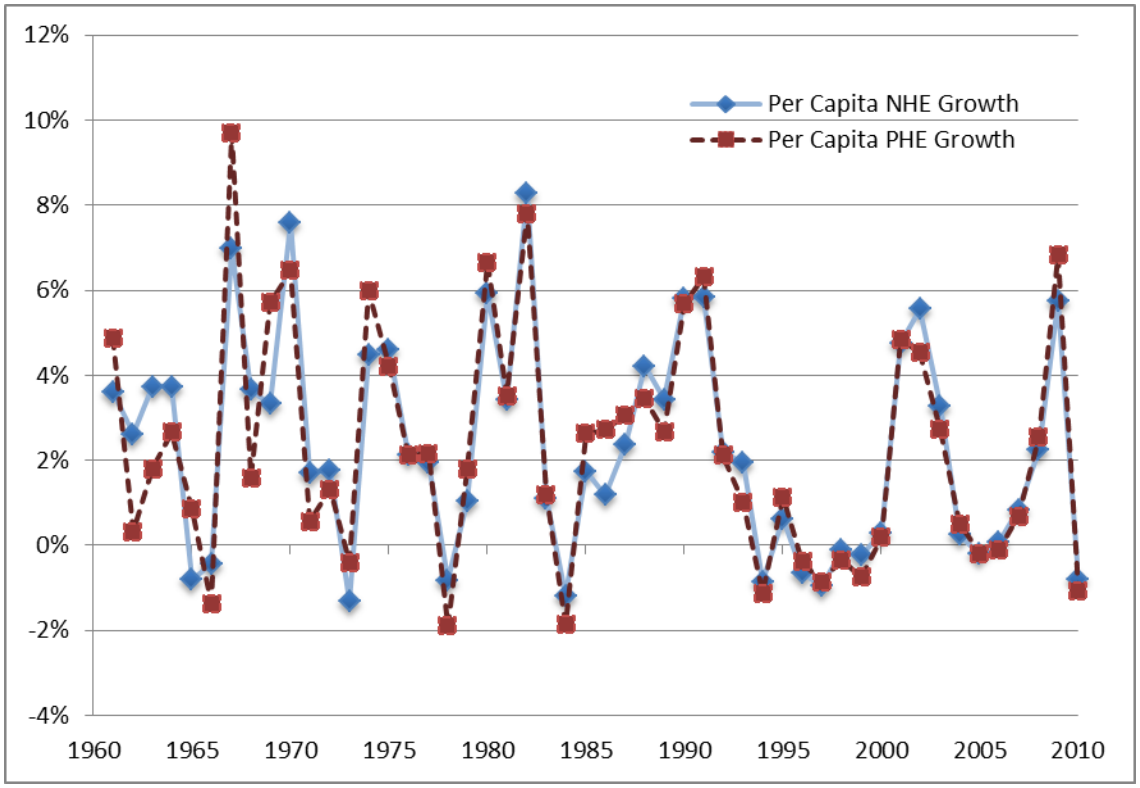
Figure 4.1: Historical Per Capita NHE and GDP



Some of the increases in per capita expenditure can be explained by demographic changes, and healthcare expenditure variables are often adjusted by an *age-gender factor* to neutralize this source. Figure 4.2 presents the annual growth rates for age-gender-adjusted per capita expenditures. In addition to the NHE series, this figure also shows corresponding growth rates for the time series computed from Personal Health Care Expenditure (PHE). Inspection of

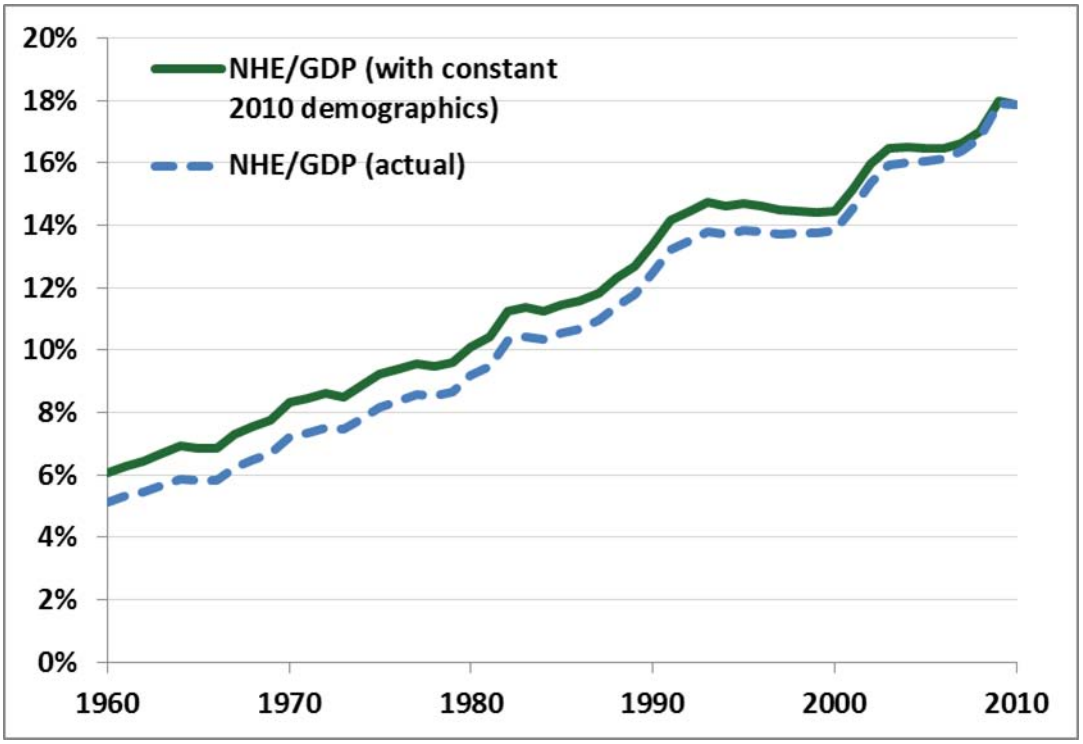
Figure 4.2 reveals that the two growth rates have quite similar profiles, which suggests that forecasts of one series fundamentally predicts the other. The aging factor used to construct these series relies on standard methods that combine 1960-1975 Census data with 1975-2010 SSA Population Data; this is the same method used by OACT and CBO.

Figure 4.2: Age-Adjusted Growth Rates in Per-Capita NHE and PHE Expenditure



While an aging population influences healthcare expenditure, it only explains a small fraction of the overall growth in expenditure. This is evident in Figure 4.3 that presents the actual NHE to GDP ratio (dotted line) along with age-gender-adjusted NHE to GDP ratio (solid line). The latter is computed as if the demographic composition were constant over the 50 year period (fixed at the 2010 composition). The actual NHE to GDP ratio has grown from 5% in 1960 to almost 18% in 2010. Had the age-gender factor instead been constant over this period (again fixed at the 2010 composition), the hypothetical NHE to GDP ratio would instead have been about 6% in 1960. So, only about one percentage point out of 13 percentage points can be attributed to the aging population over this half a century.

Figure 4.3: Historical Ratios of NHE to GDP



To formulate forecasts of healthcare expenditures, our analysis relies on data representing a measure of the excess cost ratio adjusted for age composition. Define the age-gender adjusted ratio of NHE to GDP in year *t* as

$$R_t = \frac{NHE_t}{GDP_t \cdot AgeGender_t},$$

where the age-gender factor $AgeGender_t$ is constructed by the standard methods using 1960-1975 Census data and 1975-2010 SSA Population Data. Our empirical analysis constructs data on differences over time in the natural log of R_t which equals

$$\Delta \ln R_t = \ln R_t - \ln R_{t-1} = \ln \frac{R_t}{R_{t-1}} = \ln\left(1 + \frac{R_t - R_{t-1}}{R_{t-1}}\right) = \ln\left(1 + \frac{\Delta R_{t-1}}{R_{t-1}}\right) \approx \frac{\Delta R_{t-1}}{R_{t-1}}.$$

This quantity measures growth in the ratio of NHE to GDP, which corresponds to the ECG. Our series imposes the normalization $R_{2010} = \frac{NHE_{2010}}{GDP_{2010}}$, which implies predictions of NHE/GDP hold *AgeGender* constant at the 2010 level.

4.2 Forecasts Based on a Stationary and Nonstationary AR Specification

The analysis estimates two familiar formulations of autoregressive models: (i) a stationary AR(2) model with a deterministic trend, and (ii) an AR(2) model with a unit root and

an unrestricted constant. Specification (ii) represents an integrated times series model that gives rise to a stochastic trend in the forecasts and nonstationarity. Specification (i) incorporates a linear deterministic trend in the series but is otherwise stable and corresponds in this sense to a natural analogue to the stationary specification.

Estimation of the trend-stationary AR(2) model yields the fitted specification:

$$\ln R_t = \underset{(0.14)}{1.16} \ln R_{t-1} - \underset{(0.14)}{0.29} \ln R_{t-2} - \underset{(0.18)}{0.34} + \underset{(0.0014)}{0.002667}\,(t-1960) + \hat{\varepsilon}_t \tag{4.1}$$

with $\hat{\sigma}_\varepsilon = 0.02243$; standard errors appear below corresponding coefficient estimates. Estimation of the integrated AR(2) model produces:

$$\Delta \ln R_t = \underset{(0.14)}{0.26}\, \Delta \ln R_{t-1} + \underset{(0.0045)}{0.0156} + \hat{\varepsilon}_t \tag{4.2}$$

with $\hat{\sigma}_\varepsilon = 0.02319$. Fitted specification (4.2) can also be expressed as

$$\ln R_t = 1.26 \ln R_{t-1} - 0.26 \ln R_{t-2} + 0.0156 + \hat{\varepsilon}_t\,,$$

which reveals that the autoregressive coefficients are in line with those of the trend-stationary AR(2) model. The estimated deterministic drift terms are also in the ballpark. For the trend-stationary models $\mathrm{E}(\Delta \ln R_t) = 1.16\mathrm{E}(\Delta \ln R_{t-1}) - 0.29\mathrm{E}(\Delta \ln R_{t-2}) + 0.002667$, which establishes that $\mathrm{E}(\Delta \ln R_t) = 2.0\%$ under stationarity. Correspondingly, for the integrated model, $E(\Delta \ln R_t) = 0.0156/(1\text{-}0.26) = 2.1\%$. Thus, the implied growth rates are quite similar for the two specifications.

It is straightforward to construct forecasts for future values of R_t given values for R_{2010} and R_{2009}. For the trend-stationary model this is done using the recursion

$$\ln \hat{R}_{T+h,T} = 1.16 \ln \hat{R}_{T+h-1,T} - 0.29 \ln \hat{R}_{T+h-2,T} - 0.34 + 0.002667\,(T+h-1960) \tag{4.3}$$

where $T = 2010$, and with the observed ratios $\ln \hat{R}_{2010,2010} = \ln R_{2010}$ and $\ln \hat{R}_{2009,2010} = \ln R_{2009}$ as the initial values for the recursion.

Relationship (3.2) based on the MA representation of a time series implies that the forecast errors are given by

$$\ln \hat{R}_{T+h,T} - \ln R_{T+h} = \sum_{i=0}^{h-1} \psi_i \varepsilon_{T+h-i}\ . \tag{4.4}$$

One can compute the coefficients recursively using the formula:

$$\psi_i = \hat{\varphi}_1 \psi_{i-1} + \hat{\varphi}_2 \psi_{i-2}, \qquad \text{with} \quad \psi_0 = 1 \quad \text{and} \quad \psi_{-1} = 0.$$

This yields the following estimate for the forecast error variance:

$$\hat{\sigma}_\varepsilon^2 \sum_{i=0}^{h-1} \psi_i^2 .$$

Using this quantity to form error bands yields

$$\ln \hat{R}_{T+h,T} \pm 2\hat{\sigma}_\varepsilon \sqrt{\sum_{i=0}^{h-1} \psi_i^2} . \tag{4.5}$$

The forecasts for the integrated model are computed analogously, though an important difference is that the unit root causes $\sum_{i=0}^{h-1} \psi_i^2 \to \infty$ as $h \to \infty$.

Figure 4.4 plots the projections, $\hat{R}_{T+h,T}$, implied by the two models and their corresponding error bands for this ratio, which are given by $\exp\left\{\ln \hat{R}_{T+h,T} \pm 2\hat{\sigma}_\varepsilon \sqrt{\sum_{i=0}^{h-1} \psi_i^2}\right\}$.

Figure 4.4: Forecasts of NHE as a Percentage of GDP Based on a Trend-Stationary and an Integrated Autoregressive Model

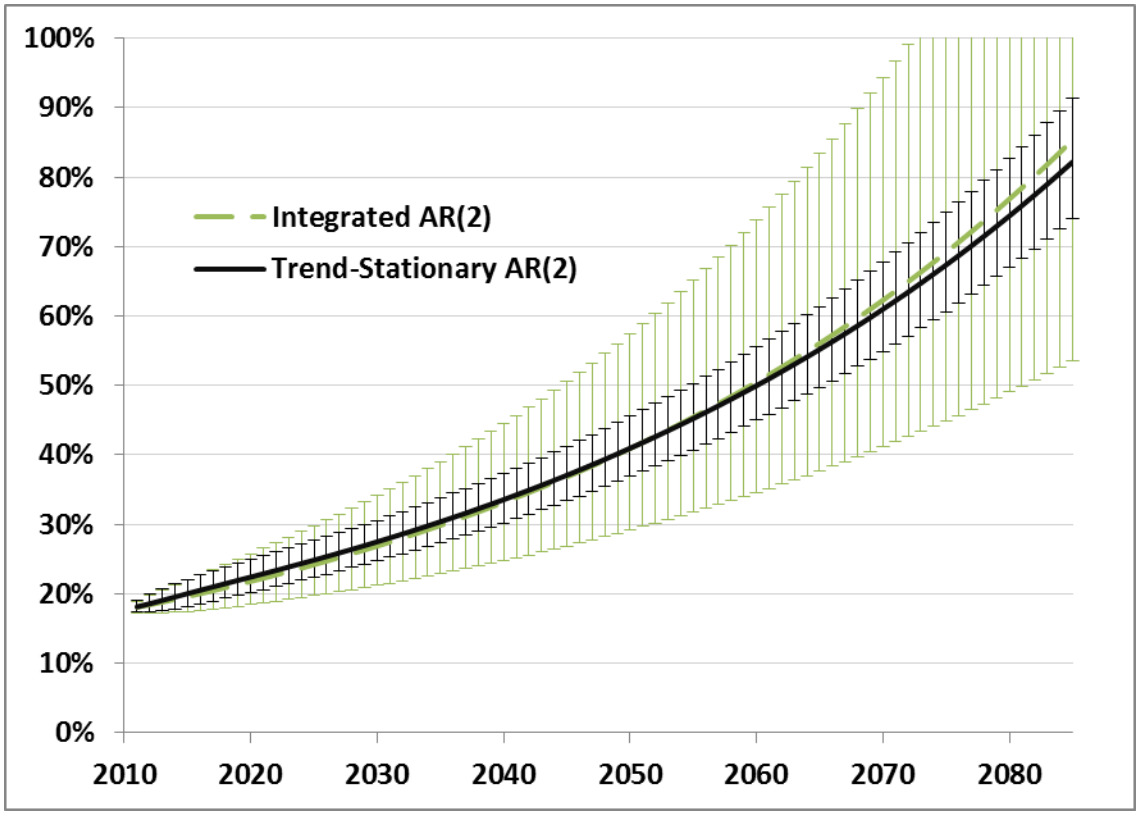


Figure 4.4 reveals that neither specification produces plausible long-horizon forecasts. The point forecasts are similar, as the forward projections are mainly driven by the linear deterministic trend. Seventy-five years out, this trend results in NHE to GDP projections that exceed 80%, and the prediction error bands exceed 100% in the case of the unit root model. These obviously unrealistic values arise because these reduced-form style models do not incorporate structural elements. The time series specifications simply assume that past trends

will continue into the distant future, without regard to natural constraints. To predict that the NHE to GDP ratio would be 18% in 2010, may have been deemed unlikely if made five decades earlier, when the NHE to GDP ratio was about 5%. While difficult to compare, it does seem rather unrealistic that the GDP share for medical expenditures would ever approach 80% given that all personal consumption which includes medical care spending now represents only 70% of GDP.

4.3 Forecasts Based on a Fractionally-Integrated AR Specification

The formulation of an Autoregressive Fractionally Integrated Moving Average (ARFIMA) model estimated in this analysis includes a constant and a linear trend, with the analytical form given by

$$(1-\varphi L)(1-L)^{d}\log R_t = a + bt + \varepsilon_t \ . \tag{4.6}$$

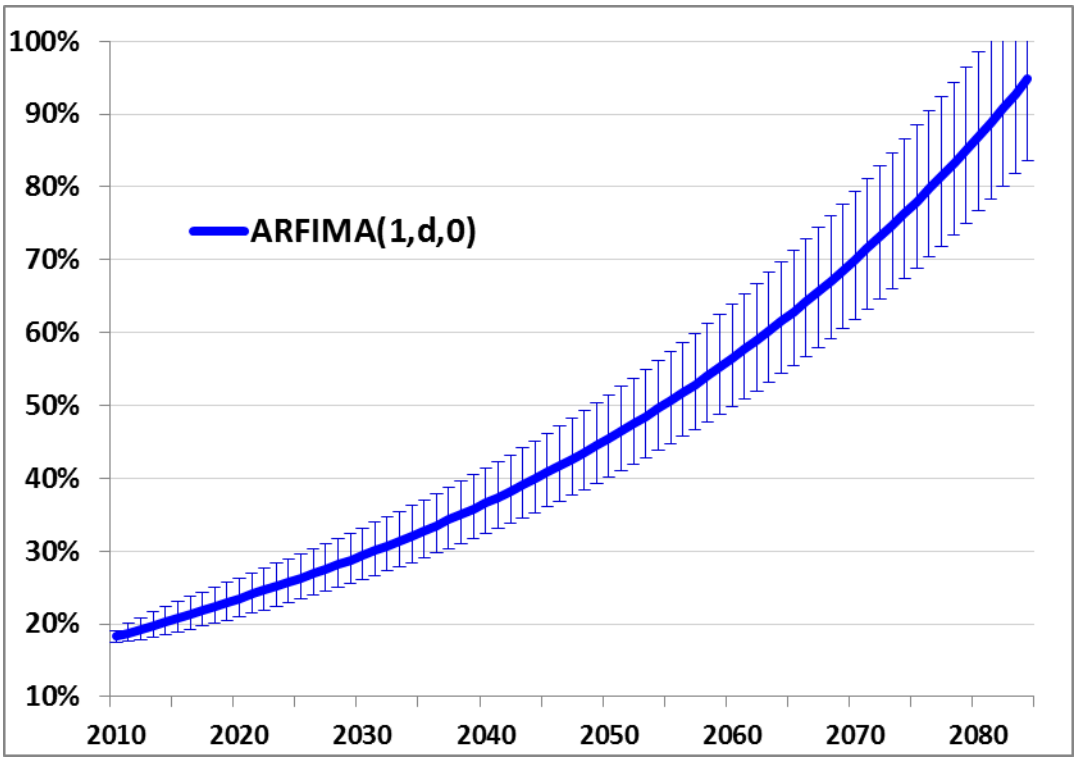
This ARFIMA(1,d,0) model is estimated by maximum likelihood using the ARFIMA package for OxMetrics, resulting in the following point estimates and corresponding standard errors in brackets:

$$\underset{(0.1481)}{\hat{\varphi} = 0.7584} \qquad \underset{(0.1938)}{\hat{d} = 0.3043} \qquad \underset{(0.060)}{\hat{a} = -0.5807} \qquad \underset{(0.00153)}{\hat{b} = 0.00216} \ .$$

These estimates imply a stationary variant of a long-memory model since $\hat{d} < \frac{1}{2}$.

Figure 4.5 presents the forecasts and standard errors produced by the ARFIMA package. The error bands are generated in the same manner as those in Figure 4.4, with confidence intervals computed as two times the standard deviation to either side.

Figure 4.5: Forecasts of NHE as a Percentage of GDP Based on a Fractionally-Integrated Autoregressive Model



The projections and error bands in Figure 4.5 are similar to those of the two autoregressive models described above. The results obtained with the fractionally integrated model are also not encouraging. The trend is estimated to be slightly larger in this model, which results in a predicted NHE to GDP ratio that is slightly higher at long horizons. In a fractionally integrated model, a shock has a lasting effect that exceeds that of stationary autoregressive models, but less than of a unit root model. This feature can be seen from the prediction error bands of the fractionally integrated models, which are wider than that of the stationary model but narrower than those of the unit root model.

4.4 Summary of Findings

Using 50 years of data on the age-gender-adjusted NHE to GDP ratio, all three fitted time-series models produce point forecasts that are simply implausible in the long horizon. When making forecasts based on reduced-form time series models, an implicit assumption is that the future will evolve in a similar manner to that observed in the past. Over the past five decades, the NHE to GDP ratio has tripled, and all the estimated time series specifications extrapolate this trend into the future. The estimated linear trend in NHE to GDP is similar for the stationary and integrated time series specifications, causing their forecasts to be similar. The different manner in which errors are accumulated in the two autoregressive specifications is

evident from the prediction error bands, as those of the unit root model continue to grow in width, eventually covering a range for which NHE exceeds GDP. The results obtained with the fractionally integrated model are also not encouraging. The trend is estimated to be slightly larger in this model, which results in a predicted NHE to GDP ratio that is slightly higher at long horizons. In a fractionally integrated model a shock has a lasting effect that exceeds that of stationary autoregressive models, but less than that of a unit root model. This feature can be seen from the prediction error bands of the fractionally integrated models which are wider that of the stationary model, but narrower than those of the unit root model.

5 CONCLUDING DISCUSSION

From a conceptual perspective, this report explains why conventional time series models fail to offer a promising framework for producing long-term forecasts, such as those in the advanced years of the 75-year horizon used to evaluate the financial viability of future healthcare spending in the US. Candidate time series specifications present an unattractive tradeoff. Whereas stationary models produce constant bands for confidence intervals, they are generally uninformative about long-horizon forecasts since they merely set these forecasts equal to an unconditional mean that does not distinguish recent from earlier historical shocks when constructing projections. Nonstationary time series models, on the other hand, allow for current trend deviations to affect future projection values, but they produce confidence intervals that explode as the projection horizon increases. Thus, forecasters selecting time series processes to make their long-term projection either encounter a situation where their estimates are precise but uninformative, or informative but imprecise.

The intuition for the inability of time series models to construct accurate long-horizon forecasts is the following: A time series model specifies the dynamic properties of a data-generating process, and these dynamic properties dictate how news about the process today will impact the process in the future. While the basic structure of a time series model can be used to make predictions at any forecasting horizon, this framework has paradoxical implications for long-horizon forecasting. For a long-horizon forecast to be informative, it is necessary that today's news matter for the future. By this we mean that the long-horizon forecast depends on the information set available today. However, if information available today is important for the distant future, then future information (that becomes available after the forecast is made) should also be important for the distant future. In fact, this structure is indeed implied by time series models. The implication of future information being important is that a long-horizon forecast is fraught with uncertainty. This *Catch 22* is the major stumbling block for long-horizon forecasting using standard time series models. After considering a broad class of time series models, including stationary and nonstationary ARMA models and short- and long-memory specifications, our conclusion is that standard time series models are—for all practical purposes—useless for making long-horizon forecasts.

While this Catch 22 tradeoff can be seen in the empirical findings in this report describing projections of future values of excess cost growth in healthcare spending, this is not the largest problem seen in the estimated forecasts. Instead, all three specifications estimated in the analysis produce point forecasts that are simply implausible in the long horizon. The analysis estimates three variants of autoregressive models covering the spectrum of candidate specifications available in the literature. The NHE to GDP ratio has tripled in the past 50 years, and all estimated specifications extrapolate this trend into the future. While this rate might continue in the next decade or so, it cannot continue for another 50 to 75 years. The stationary

model with a linear trend produces essentially the same point estimates as the nonstationary model with a unit root. The results obtained for the fractionally integrated model imply slightly higher forecasts. All estimated specifications forecast NHE as a percentage of GDP to increase over 10 percentage points in two decades—from just under 20% now to over 30% by 2030—and exceed over 80% by the end of the 75-year forecast horizon.

The underlying trends are key for forecasting healthcare costs several decades into the future. A major obstacle is that these trends cannot be estimated accurately from past trends, for the simple reason that past trends have been too large to be sustainable for much longer. Standard time series models produce naïve predictions of trends, with prediction bands that are typically either too large or too small depending on whether a researcher adopts a nonstationary or stationary specification. A potential modification for overcoming these disadvantages could be to formulate forecasts incorporating ancillary variables—such as income and medical practice in the case of health expenditures—to reflect that as healthcare expenditures grow relative to GDP, economic and political forces will eventually slow this growth. Because no evidence of such forces currently appear in the historical data, reduced-form time series specifications fail to account of such factors. Instead, economic models or other *a priori* structural relationships would be required to provide the principal foundation for specifying long-run trends and far-distant forecasts. Although time series models could be adapted to incorporate such components, the resulting frameworks would go well beyond familiar specifications. In such instances, conventional time series models could be used to describe short-term fluctuations around trends, but they would likely play only a minor role in constructing long-term forecasts for the reasons discussed in this paper.

REFERENCES

Beveridge, S. and C. R. Nelson (1981). "A New Approach to Decompositions of Time Series Into Permanent and Transitory Components with Particular Attentions to Measurement of the Business Cycle". Journal of Monetary Economics, vol. 7, 151-174.

Borger C, Rutherford TF, Won GY. Projecting long-term medical spending growth. Journal of Health Economics. 2008;27(1):69–88.

Box, G. E. P. and G. M. Jenkins (1976). Time Series Analysis; Forecasting and Control. Prentice Hall.

Caldis, TG and U.S. Department of Health and Human Services. The long-term projection assumptions for Medicare and aggregate national health expenditures. Washington, DC: Centers for Medicare & Medicaid Services; 2009.

Clements, M. P. and D. F. Hendry (1998), Forecasting Economic Time Series, Cambridge University Press, Cambridge.

Congressional Budget Office (2012). The 2012 Long-Term Budget Outlook. June 2012. http://www.cbo.gov/sites/default/files/cbofiles/attachments/06-05-Long-Term_Budget_Outlook_2.pdf

Congressional Budget Office (2012). Supplemental Data to the 2012 Long-Term Budget Outlook. Accessed Dec 10, 2012. http://www.cbo.gov\sites\default\files\cbofiles\attachments\43288-LTBOSuppTables_0.xls

Diebold, Francis X. (2006). Elements of Forecasting. 4th ed. South-Western College Pub.

Federal Old-age and Survivors Insurance and Federal Disability Insurance Trust Funds (OASDI) Board of Trustees (2012). The 2012 Annual Report of the Board of Trustees of the Federal Old-age and Survivors Insurance and Federal Disability Insurance Trust Funds . Washington, April 25, 2012. http://www.ssa.gov/oact/tr/2012/tr2012.pdf

Friedman, JN and National Research Council (US) Committee on National Statistics. (2010), "Improving Health Care Cost Projections for the Medicare Population: Summary of a Workshop.", Washington (DC): National Academies Press (US); Appendix A, Predicting Medicare Cost Growth. Available from: http://www.ncbi.nlm.nih.gov/books/NBK52815/

Getzen T. Long-Run Forecasts of National Health Expenditure Growth. SSRN Working Paper, October 28, 2011. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1950810

Granger, C. W. J. and P. Newbold (1986). Forecasting Economic Time Series. 2nd ed. Academic Press.

Hall RE, Jones CI. (2007), "The value of life and the rise in health spending." Quarterly Journal of Economics;122(1):39–72.

Harris KM, Galasso JP, Eibner C, editors. U.S. Department of Veterans Affairs. (2008), "Review and evaluation of the VA enrollee health care projection model." RAND Corporation for the Department of Veterans Affairs.

Hamilton, James D. (1994). Time Series Analysis. Princeton N.J.: Princeton University Press.

Hansen, B. E. (2007). "Least squares model averaging", Econometrica vol. 75, pp. 1175-1189.

Hansen, P. R. (2005). "Granger's Representation Theorem: A Closed-Form Expression for I(1) Processes". Econometrics Journal, vol. 8, pp.23-38.

Heffler, Steve, T. Caldis, and S. Smith. (2012) "The Long-Term Projection Assumptions for Medicare and Aggregate National Health Expenditures." Centers for Medicare and Medicaid Services Memorandum, November 27, 2012. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ReportsTrustFunds/Downloads/ProjectionMethodology2012.pdf

Johansen, Søren (1988). "Statistical Analysis of Cointegration Vectors". Journal of Economic Dynamics and Control, vol. 12, pp. 231-254.

Johansen, Søren and Morten Ø. Nielsen (2009). "Likelihood Inference for a Nonstationary Fractional Autoregressive Model". Unpublished manuscript.

Lee, R. and Miller, T. (2002), An Approach to Forecasting Health Expenditures, with Application to the U.S. Medicare System. Health Services Research, 37: 1365–1386. doi: 10.1111/1475-6773.01112

National Health Expenditure Accounts (NHEA). NHE summary including share of GDP, CY 1960-2010. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/NHEGDP10.zip

Office of the Chief Actuary of the Social Security Administration (2012). The Long-Range Demographic Assumptions for the 2012 Trustees Report. April 23, 2012. http://www.ssa.gov/oact/tr/2012/2012_Long-Range_Demographic_Assumptions.pdf

Shatto, JD and Clemens MK. (2012) "Projected Medicare Expenditures under Illustrative Scenarios with Alternative Payment Updates to Medicare Providers." Centers for Medicare and Medicaid Services Memorandum, May 18, 2012. http://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/reportstrustfunds/downloads/2012tralternativescenario.pdf

Stock, J. H. and Watson, M. W. (2002). "Forecasting using principal components from a large number of predictors". Journal of the American Statistical Association vol. 97, pp.1167-1179.