2007-10-26

# Accounting for Additional Heterogeneity: A Theoretic Extension of an Extant Economic Model

Bradley John Barney
*Brigham Young University - Provo*

# ACCOUNTING FOR ADDITIONAL HETEROGENEITY: A THEORETIC EXTENSION OF AN EXTANT ECONOMIC MODEL

by

Bradley J. Barney

A Project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

December 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a Project submitted by

Bradley J. Barney

This Project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

| | |
|---|---|
| Date | H. Dennis Tolley, Chair |
| Date | Bruce J. Collings |
| Date | David G. Whiting |

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the Project of Bradley J. Barney in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

H. Dennis Tolley
Chair, Graduate Committee

Accepted for the Department

Scott D. Grimshaw
Graduate Coordinator

Accepted for the College

Thomas W. Sederberg
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

ACCOUNTING FOR ADDITIONAL HETEROGENEITY: A THEORETIC EXTENSION OF AN EXTANT ECONOMIC MODEL

Bradley J. Barney

Department of Statistics

Master of Science

The assumption in economics of a representative agent is often made. However, it is a very rigid assumption. Hall and Jones (2004b) presented an economic model that essentially provided for a representative agent for each age group in determining the group's health level function. Our work seeks to extend their theoretical version of the model by allowing for two representative agents for each age—one for each of "Healthy" and "Sick" risk-factor groups—to allow for additional heterogeneity in the populace. The approach to include even more risk-factor groups is also briefly discussed. While our "extended" theoretical model is not applied directly to relevant data, several techniques that could be applicable were the relevant data to be obtained are demonstrated on other data sets. This includes examples of using linear classification, fitting baseline-category logit models, and running the genetic algorithm.

# ACKNOWLEDGEMENTS

Many people assisted me with this work. I would like to thank in particular Dr. Tolley and my other committee members, as well as my fellow students and various members of the faculty who gave me specific direction in various aspects. Dr. Mark Showalter of the Department of Economics also provided valuable assistance in the early stages of this work. Last of all, I express gratitude to my wife, Tammy, for her encouragement.

CONTENTS

# TABLES

Table

# FIGURES

Figure

# 1. INTRODUCTION

Those involved in the social sciences have a vested interest in explaining why people make the choices they do. One important result of understanding a person's decision-making processes is the insight that can be gained into deducing that person's likely responses in various situations. As an economics example, knowing what the likely effect of a consumption tax on cigarettes will be on cigarette use can help in making decisions about whether or not to impose the tax, and if so, how high it should be to attain some desired result.

One point worthy of mention before proceeding is the assumption in economics that there is a representative agent who can characterize aggregate results. This assumption is often made to avoid staggering complexities which would be present if the behaviors of numerous unique agents were to be dynamically and simultaneously modeled. In effect, what is assumed is that the aggregate response in a given situation can be expressed by using only one agent. Frequently, such a representative agent is assumed to behave a certain way in order to optimize some specified function, usually in the presence of specified constraints. As an example, a utility function with the variables consumption and leisure time might be given, and the representative agent's behavior (given time and income restrictions) which produces the greatest possible utility for that person would be assumed to characterize the aggregate behavior of all agents under consideration by assuming that the agent's actions are repeated once for each member of the group the agent represents. This assumption greatly simplifies the task of modeling large-scale results in an economy, but it is not always an ideal assumption, as is discussed in Section 2.2.

Identifying a way to move away from the strict assumption of having a characteristic agent in order to include more heterogeneity is a worthwhile endeavor in

creating more realistic models. While there are different possible means of implementing more variety in macroeconomics, the method proposed here looks at a specific example of how more heterogeneity can be included in an economic model.

In the earlier stages of this work, Hall and Jones had a copy of one of their papers available online, referred to hereafter as Hall and Jones (2004b), which has since been published in a modified form. The recently published version of this paper is referred to herafter as Hall and Jones (2007). Some of what is contained in this work does not apply to the (2007) version of their paper, which is undesirable because it is their final version. However, because much of the work in this document was done before the recent version of the Hall and Jones article was published, this work is based on the (2004b) version of the Hall and Jones article. Because this version is no longer readily available online, both versions are cited when there is agreement between the two. If the articles do not agree, a version preceding the other two articles, Hall and Jones (2004a), is cited along with the (2004b) paper. However, this paper will refer to the model as the Hall and Jones (2004b) model, but it is still quite similar to the (2007) model. Note also that understanding of Hall and Jones (2004b) was deepened by Dr. Mark Showalter in personal communications from May–July 2005 and by discussion with Dr. Dennis Tolley throughout 2005; these comments have influenced the provided summary of Hall and Jones (2004b).

Hall and Jones (2004b) undertook, among other things, to provide insight into the fact that United States consumers are spending a greater percentage of their earnings on health care now than they did previously; the percentage of earnings spent nearly tripled from 1950 to 2000 (Hall and Jones 2004b, p. 2; Hall and Jones 2007, p. 39). Through their work, they show that substantial growth in the proportion of total income spent on health is in concordance with optimizing behavior in a fictional economy with specified properties (M. Showalter, personal communications, May–July 2005). Hall and Jones initially present a very rigid economic structure but then

relax several of their assumptions and apply their model to economic and health data from the United States. They rely on the assumption that a representative agent can be used to characterize all people in the economy of the same age at a given time throughout their paper.

The aim of the research proposed here is to show how a model can be modified to contain additional heterogeneity. For illustration purposes, the Hall and Jones (2004b) model will be modified in this way. Explicitly, Hall and Jones (2004b) assume that age is the only difference among individuals' health functions at a given time; thus, a representative agent for each age group is sufficient to model behavior in the health economy at any given time (see also Hall and Jones 2007). In this project their model is extended to include risk factors. Recall that the purpose of this project is to illustrate how to adapt a model to allow for more variety in the populace than the model would otherwise allow. There is a substantial increase in complexity associated with a very general result. The principles used in extending the Hall and Jones (2004b) model are illustrated with only two risk-factor states; that is, it is assumed there are two representative agents for each age group and that knowing the behaviors of these representative agents is sufficient to determine aggregate behavior. The two risk-factor groups, which have age-specific properties, are generically termed "Healthy" and "Sick."

The extended model presented in this project will make a weaker assumption than the representative agent assumption used by Hall and Jones (2004a,b)—it will assume that people are exactly the same with respect to the model in a given time period if they are the same age and belong to the same risk-factor group. Strictly speaking, the extended model presented can be considered to be a model with three risk-factor groups, with the third group being "Deceased," but this group will be treated fundamentally differently. People in the "Deceased" risk-factor group are not modeled further once they enter this group, because the purpose of this group is to

allow for an individual to always be in a risk-factor group in future periods. Thus, there is no associated representative agent for people in this risk group.

Not only will there be two representative agents for each age group, one for those in the "Healthy" group and one for those in the "Sick" group, but agents in either of these risk-factor groups in one discrete time period could stay in that group, switch to the other group, or be in the "Deceased" group in the next time period. The ability of individuals to switch between risk-factor groups is an important feature of the proposed model because it incorporates more heterogeneity than simply having different groups. It will be assumed that individuals in the "Deceased" risk-factor group cannot later enter the "Healthy" or "Sick" groups.

While the extension of including these risk-factor groups might seem very trivial, there is a significant amount of change involved in doing so. It is also important to recognize that this is a first step. Once the model has been extended to two risk-factor groups, it is much easier to extend the model to more groups. While it will be noted how to adapt the model for more risk-factor groups, in almost all instances, the presentation involves only the risk-factor groups "Healthy," "Sick," and "Deceased."

It is important to recognize that the goal of this paper is to illustrate a procedure by which additional heterogeneity can be accounted for in an economic model, as summarized above and as detailed throughout the rest of this document. While some of the data necessary to actually apply the extended model to the U.S. economy are currently being gathered, all of the necessary data to fit the theoretical model are not available. For the reader interested in understanding how certain practical techniques for fitting the extended model could be implemented, some examples requiring data are given based on another type of currently available data or using assumed values in place of unknown quantities. It is important to note that the data used are not claimed to be related to the extended version of Hall and Jones' (2004b) model, but these examples are useful as an example of how the appropriate data might be handled

if it were available.

A close-to-optimal numeric solution of the extended model only makes sense if necessary model parameters are well estimated or reasonably assumed. This project does not claim to use reasonable values in all instances, and the data to estimate all of the parameters is not available. A procedure for obtaining a close-to-optimal numerical solution for maximizing a particular function will be demonstrated for a very simplified version of the problem using assumed values for all unknown model parameters. Although this makes the final results meaningless, it provides an example of the numerical solution tool to be used. It cannot be overstated that this paper attempts to inform the reader how increased heterogeneity could be accounted for in economic modeling by extending a posed theoretical model by Hall and Jones (2004b) and does not attempt to sustain nor reject Hall and Jones' findings from having fitted their model with U.S. data.

# 2. LITERATURE REVIEW

## 2.1 Hall and Jones (2004b) Model

While there is a great deal of information presented in Hall and Jones (2004b), the focus of this project is on the model itself, as opposed to the implications of the fitted model or the details of how Hall and Jones fitted their model, given that alternative methods are sometimes suggested herein. Note that the Hall and Jones (2004b) model will be closely followed in the "extended" version of this model.

Essentially, in the principal model of Hall and Jones (2004b), they assume a deterministic function for the health level, $x_{a,t}$, of any individual of age $a$ at time $t$ as an age-dependent function of that individual's "effective health input," $z_t h_{a,t}$, where $z_t$ represents the health technology level at time $t$ and $h_{a,t}$ denotes the individual's health spending. This health level is also used as the inverse of the probability that a corresponding individual will live to the next period. Hall and Jones also assume a particular form for the utility of a person of age $a$ at time $t$ as a function of the individual's health level and the individual's consumption spending for that period, $c_{a,t}$. They pretend that someone in charge of the economy wants to maximize a function consisting of the sum of all individuals' utility in this period and a diminishing proportion of the total utility in every future period (Hall and Jones 2004b, pp. 13–14; see also Hall and Jones 2004a, pp. 15–16).

There are a number of constraints placed on the problem, including total income equals total health spending plus total consumption spending in each period; the same number of people are born each year, denoted by $N_{a=0,t} = N_0$; the health levels determine the proportion of individuals of each age that will be alive in the following time period; health technology level has exponential growth; income is the same across individuals, and this common level of per-capita income, $y_t$, also has

exponential growth; the health function determines health level. After posing this theoretical model, they impose a particular form on the health functions and then apply this model to U.S. data under a pair of scenarios to discuss the past *and* make future projections (Hall and Jones 2004b, pp. 2–3, 14–15, 18; see also Hall and Jones 2004a, pp. 2–3, 16–17, 20).

Again, Hall and Jones did much more, but this review represents an overview of what is needful to understand the extended model. Specific information will be provided throughout this paper. While there are admittedly important features of Hall and Jones' paper that have not been discussed, they will not be addressed herein. The focus of this paper is simply on how a certain aspect of their theoretical model could be extended to include more representative age groups and to allow switching between risk-factor groups. Also, means to apply this model in the future provided someone were to collect the relevant data is generally described, but no attempt is made to fit the extended theoretical model to past data, and estimation of some important parameters is not addressed at all in the extended model.

## 2.2 Representative Agent Literature

The assumption of one representative agent characterizing all members of a macroeconomy is rather common. Highlights of the development of using representative agents are briefly noted in Martel (1996), and Hartley (1996) provides a much more in-depth account. The concept of a representative entity being used to model the economy was initially created by Marshall, who in 1920 introduced the concept of a representative firm (Hartley 1996, p. 169; Martel 1996, p. 128). However, his usage of the term *representative* had a quite different implication than it does now (Hartley 1996, p. 169). According to Hartley, Marshall was seeking to find a way to get a unique price and equilibrium at the aggregate level while still allowing firm-to-firm differences (Hartley 1996, p. 171). This can be compared with Hall and Jones'

(2004b) usage of a representative agent for each age group, wherein they assume that all persons of the same age at a given time are exactly the same in their health functions (Hall and Jones 2004b, pp. 7, 13; see also Hall and Jones 2007, pp. 44–45, 49).

Despite the rather widespread use of the representative agent supposition, there are criticisms against it. Besides mentioning other arguments, Kirman (1992) makes the following four arguments against using representative agents:

(1) No strong support exists that collective actions of individual optimizers would behave as just one optimizing person would.

(2) Discounting the first argument still leaves the flaw that altering a model parameter might produce a discrepancy between what the representative agent would do and what the collective actions would be (in the presence of the alteration).

(3) It is possible that every individual forming the aggregate would prefer the second of two possibilities, but the representative agent would prefer the first possibility.

(4) Relying on a representative agent in empirical work has the drawback of being very restrictive insofar as collective actions might be very complex but only one individual can be used to characterize them.

Similarly, Carroll (2000) presents a result which he claims is cause to stop using the representative agent approach in certain cases and to instead use an approach that is in line with "key microeconomic facts." This claim follows from his deduction that sometimes an altered representative agent approach (altered by including idiosyncratic risk) does not give outcomes consistent with these microeconomic facts.

He finds that in an example where he tries to adapt a model to eliminate these inconsistencies, the results are substantially different from those obtained using the representative agent approach (Carroll 2000, pp. 110–114).

Many other economists have also seen the weaknesses of using representative agents. Because of this, methods of altering or bypassing this assumption have arisen. Martel (1996) provides a discussion on problems with using representative agents, as well as a summary of several paths alternative research is taking. He mentions two areas which are still related to the representative-agent case. The first area he describes as dividing the macroeconomy into different non-aggregated sectors, each of which is then assumed to have a representative agent. He describes a second type of research that modifies the representative agent assumption as the use of extra variables to incorporate characteristics of cross-section-specific distributions (Martel 1996, pp. 137–138).

Daniel's (1993) paper provides an example of research with more than one representative agent. In her work, Daniel explores the relationship between the timing of taxes (i.e., in what time period(s) taxes are levied) and the macroeconomy. She uses two representative families instead of just one, as this allows for families with either of two discount parameters. She finds that using two representative families instead of just one substantially alters the implied macroeconomic impact caused by the timing of taxes.

Similarly, Hall and Jones (2004a,b) used a different representative agent for each age group rather than just one overall. The extended model is based on having multiple representative agents for each age group. Thus, this is the technique used in the extended model to avoid the restriction of having only one representative agent overall and to allow for more heterogeneity in the model.

Caselli and Ventura (2000) use the assumption that there is a representative consumer who can characterize what all consumers do on average. They note that

to be consistent with this assumption, heterogeneity among consumers must be incorporated in an acceptable manner, but they point out that the assumption of a representative consumer does not preclude heterogeneity of consumers. Caselli and Ventura discuss the inclusion of heterogeneity in three consumer traits and give examples of models that have heterogeneous consumers and yet are consistent with the assumed existence of a representative consumer. They argue that the assumption of a representative consumer causes relatively little further restriction of the observable dynamics of model results beyond what the observable dynamics would be without the representative consumer assumption (Caselli and Ventura 2000, pp. 909, 911–912, 923).

Martel (1996) also notes work by Hildenbrand that, instead of trying to adapt the representative agent approach, depends on agent differences for having a stable aggregate demand (Martel 1996, p. 138). Related work leads Russell (1995) to conjecture that having enough heterogeneity present in his model could imply "approximate" holding of the Coase theorem (Russell 1995, p. 105). In recent work by Hildenbrand and Kneip (2005), they note that as a type of diversity with regard to households increases, as measured by their proposed index, the impact that prices have on "the aggregate consumption expenditure ratio" decreases, under the condition that there are not an infinite number of households so that the index proposed can be used (Hildenbrand and Kneip 2005, p. 155). Note that in each of the aforementioned cases heterogeneity seems to be a desirable trait rather than a nuisance.

## 2.3 Relevant Health Economics Literature

In addition to the Hall and Jones (2004b) paper, there is a great deal of research in health economics which is relevant to the proposed extension of the Hall and Jones model. One key assumption in Hall and Jones (2004a,b) is that a person's health level and utility depend, in part, on the person's health expenditures. Nordhaus (2003)

claims that empirical evidence and reason seem to indicate economically quantifiable health gains are at most modestly related to health spending. He notes that an alternative way to assess health contributions is to look at how many doctor visits people make or how many days they spend in a hospital (Nordhaus 2003, pp. 10–11).

Nordhaus' statement is particularly meaningful because Hall and Jones (2004b, 2007) suppose that a person's health level is a deterministic function of health spending, among other things. While this may be a less-than-ideal assumption about what determines a person's health, especially in light of Nordhaus' comment, there is an important point which hopefully reduces the anxiety involved with making this assumption. Recall that the goal of the proposed research is to show how heterogeneity might be added to an economic model. As part of the Hall and Jones (2004a,b) model, the dependence of health level on health expenditures was an integral element in the purposes of their research, and such dependence also helped to make the model more tractable than it would otherwise be. It makes sense to perpetuate a dependence of health on health expenditures in the extended model, which is discussed in Chapter 3.

Literature which seeks to assess how medical treatments affect people is also relevant. Meltzer points out that when analyzing "medical interventions" in terms of how much gain is derived from them versus how much they cost, a particular obstacle is having the means to summarize the benefits associated with an intervention using just one number. He notes that the increase in life expectancy associated with different interventions can be compared, even for different medical conditions, yet the comparison of life expectancies leaves something to be desired because it does not incorporate how the intervention impacts the enjoyment of life. He describes the advent of quality-adjusted life years (QALY's) as means to overcome this deficiency. Essentially, QALY's assign weights ranging from zero to one to each life year, where the weights correspond to how healthy the person is in that life year. These weights then provide information on how enjoyable that life year is from a health perspective

(Meltzer 2003, p. 215).

Quality-adjusted life years have become widely used, but they are not embraced by all as an ideal assessment tool (Meltzer 2003, p. 216). Cox, Fitzpatrick, Fletcher, Gore, Spiegelhalter, and Jones argue that instead of always converting a multivariate treatment impact into a univariate measure, in certain situations the multivariate treatment impact should be stated so that individuals can then make a choice suited to their desires; the treatment impact is not based on a multivariate-to-univariate projection using other individuals' preferences (Cox et al. 1992, p. 354). Meltzer argues in effect that even though multiple methods of estimating the weight parameters in QALY's have often led to correlated results, the correlation in the results can only be used to conclude that QALY's can be precisely (and possibly not validly) measured (Meltzer 2003, p. 216).

Although Hall and Jones (2004b) did not state that they were using QALY's, they included the health level in the utility function and set the inverse of the health level equal to the mortality rate (Hall and Jones 2004b, pp. 12–13; see also Hall and Jones 2004a, pp. 14–15). Thus, the health level as used in their model does impact both the enjoyability and the expected duration of life (Hall and Jones 2004b, p. 13; Hall and Jones 2007, p. 49). The extension of their model will also allow for health spending to influence the enjoyability and expected duration of life. However, a concept somewhat like that used to motivate the use of QALY's will be used in the health level to allow for the probability of being in a "Healthy" state in the next period to have a bigger weight than the probability of being in a "Sick" state. For the details of this, see the discussion in Section 3.1.3.

## 2.4 Classification Literature

### 2.4.1 Introduction to Classification Techniques

Putting "classification analysis" into a more narrow context, one use of classifying is to determine to which of multiple populations an individual belongs. To make accurate classifications, an important preliminary step is to determine the characteristics of individuals in each of the populations. Measurements taken on people who have already been correctly grouped can help in identifying these characteristics because they might provide insight into the types or ranges of measurement values that are likely to be observed in each population. When some future person has unknown population membership, that person can be measured. That person's likely membership is determined based on how his or her observed measurements compare to what was seen in each population's reference group (Rencher 2002, p. 299).

Rencher provides an exposition of several different classification techniques. Among these, he describes linear and quadratic classification techniques (Rencher 2002, chap. 9). If the covariance matrix associated with the vector of random variables to be measured on an individual from a given population is assumed to be the same across populations, the former technique can be used, but if not then the latter can be used.

Of particular interest, Rencher describes a linear classification technique used by Fisher (1936) for classifying with two candidate populations. Rencher notes that this technique has the advantage of not depending on the random variables having a multivariate-normal distribution (Rencher 2002, pp. 300–302). However, to incorporate important prior information, a similar linear classification criterion based on a rule by Welch (1939) does make the assumption of multivariate normality (Rencher 2002, p. 302). Rencher also notes that both of these linear classification techniques assume the covariance matrices to be the same for both populations (Rencher 2002,

pp. 300, 302). He also explains linear classification rules when there are more than two candidate populations, both for when there is prior information on the proportion of individuals in each population as well as for when there is not prior information (pp. 304–305).

Rencher explains the quadratic classification technique used when the covariance matrices are not assumed to be the same for both populations, but this technique requires a multivariate normal distribution for the random vector. Unless a substantial amount of data is available or the covariance matrices differ greatly from each other, using linear rather than quadratic classification might be advantageous because of the instability associated with estimating a different sample covariance matrix for each population (Rencher 1998, p. 232–233).

Rencher also describes a technique based on the groups to which similar individuals belong and a technique using a kernel density estimate, both of which are nonparametric approaches (Rencher 2002, chap. 9). While the technique using the kernel density estimate does not make strong parametric assumptions or require the covariance matrices to be the same in each population, it is very sensitive to the value of a parameter which characterizes how much the estimated density should be smoothed (Rencher 2002, pp. 315–317). It seems likely that a large number of response vectors would have to be gathered to reasonably estimate a multivariate density function with this nonparametric technique, especially as the response vector's length grows. Fix and Hodges (1951) introduced a nonparametric classification technique that is called the $k$ nearest neighbor rule. The rule uses the $k$ observations (among all the observations from individuals with known population membership) that are closest to the observation from an individual with unknown population membership. The rule then classifies the individual as belonging to the population that is most prevelantly represented by these $k$ observations. In measuring the closeness of any two observations, the rule takes into account the covariance matrix of the observations. The

rule uses a pooled estimate of the covariance matrix and is also sensitive to the value of $k$, or how many nearby observations are considered when classifying an individual (Rencher 2002, pp. 318–319). The use of a pooled covariance matrix estimate might hurt the technique's classification accuracy if the covariance matrices are wildly different from each other across populations. Likewise, the sensitivity to $k$ leaves the possibility of poor classification performance if an inappropriate value of $k$ is chosen.

If the measurements being taken on individuals are all categorical, another technique explained by Rencher (2002) might be considered (Rencher 2002, p. 314). This technique uses a rule proposed by Welch (1936) which classifies an individual into a population based on the ratio of the multivariate density of one population evaluated at the individual's response vector to the ratio of the multivariate density of another population, again evaluated at the individual's response vector (Rencher 2002, p. 314). To estimate the multivariate density of a random response vector from a given population, Rencher explains that we could use the proportions of observed response vectors (from the given population) that equal each possible response vector. A considerable drawback to this technique is the need to have large sample sizes from each population to ensure accurate estimates (Rencher 2002, pp. 314–315). The need for large sample sizes becomes more and more pressing as more response vectors become possible.

A markedly different technique, Grade of Memberships modeling (GoM), is described in detail in Manton, Woodbury, and Tolley (1994). A fundamental notion in this technique is that rather than one individual's characteristics pertaining entirely to a population, the individual can be considered as pertaining to each population with a proportion from 0 to 1 inclusive, where the sum of the proportions is one (Manton et al. 1994, chap. 1). However, in the proposed extension of the Hall and Jones (2004b) model, it will be assumed that individuals are wholly in one population at each time period, so this technique cannot be directly applied unless this assumption

were relaxed. Nonetheless, the GoM model could be used in an indirect fashion by first estimating each individual's grade of membership in each population with GoM modeling and then classifying the individual into a population based on these estimated grades (H. D. Tolley, personal communication, June–July 2005). This method might be advisable if all of the measurements being taken on an individual are categorical (H. D. Tolley, personal communication, July 2005) because the GoM technique was created specifically to deal with a response vector of many dimensions containing categorical responses (Manton and Land 2000, p. 198).

None of the techniques mentioned so far seem clearly superior in all situations for classification into one of multiple populations based on an individual's response vector. If all variables are continuous, there is still no single technique that seems clearly superior if we assume neither multivariate normality nor covariance matrices that do not differ across populations. However, the linear classification technique using prior information, which does make both of these assumptions, seems promising if there is not a great quantity of data. For instance, if there is not a sufficiently large amount of data, allowing for different covariance matrices with the alternative quadratic classification technique could be problematic because the sample covariance matrices used in place of a single pooled covariance matrix would not be stable (Rencher 1998, p. 233).

The only classification technique that will be demonstrated in this paper is the linear classification technique using prior information. The other techniques have been mentioned to inform the reader of other methods which exist for classifying individuals into populations based on response vectors.

### 2.4.2 Details of the Linear Classification Technique

The technique for classification used in this project is the version of the linear classification technique for multiple populations that incorporates prior information

(Rencher 2002, pp. 304–305). This technique assumes that a random response vector from any of the candidate populations has the same covariance structure as a random response vector from any of the other candidate populations (Rencher 2002, pp. 304–305).

The remainder of this section describes the linear classification technique with $M$ candidate populations, with some changes to Rencher's (1998, 2002) notation to aid in understanding. Let $p_m$ represent the probability prior to collecting measurements on an individual that he or she is in the $m^{\text{th}}$ population, $m = 1, \ldots, M$ (Rencher 1998, p. 230). This project assumes that the prior probability of each population is known. If the prior probability of each population is not already known, then the prior probability for a given population might be assumed to equal the proportion of individuals in a reference set that pertain to the given population. Such a reference set should be obtained by randomly sampling from the collection of all individuals and determining population membership for each individual in the sample.

We use $\mathbf{y}$ to denote the vector of measurements from an individual (Rencher 1998, p. 230). The classification rule classifies an individual into the $m^{\text{th}}$ population if

$$\ln(p_m) + \bar{\mathbf{y}}_m' \mathbf{S}_{pl}^{-1} \mathbf{y} - \frac{1}{2} \bar{\mathbf{y}}_m' \mathbf{S}_{pl}^{-1} \bar{\mathbf{y}}_m > \ln p_i + \bar{\mathbf{y}}_i' \mathbf{S}_{pl}^{-1} \mathbf{y} - \frac{1}{2} \bar{\mathbf{y}}_i' \mathbf{S}_{pl}^{-1} \bar{\mathbf{y}}_i$$

for all $i \neq m$ (Rencher 2002, p. 305). Note that the previous equation is substantially altered from the presentation in Rencher (2002). For clarification, $\bar{\mathbf{y}}_m$ represents the mean response vector for those individuals in the reference set from the $m^{\text{th}}$ population (Rencher 2002, p. 304). Also, $\mathbf{S}_{pl}$ represents the estimated covariance matrix under the assumption that each population has the same covariance matrix. The estimated covariance matrix is calculated as

$$\mathbf{S}_{pl} = \frac{1}{N - m} \sum_{m=1}^{M} (n_m - 1) \mathbf{S}_m,$$

with $n_m$ denoting how many individuals from the reference set were in the $m^{\text{th}}$ popu-

lation, $N$ denoting how many individuals were in the reference set, and $\mathbf{S_m}$ denoting the sample covariance matrix for the $m^{\text{th}}$ population (Rencher 2002, p. 304; note minor change in notation).

Given the previous assumptions in this subsection, the additional assumption that the distribution of each population's response vectors is multivariate normal yields a desirable property. This desirable property is that the aforestated rule is a sample analog of the classification rule with the smallest misclassification probability (Rencher 2002, pp. 302, 304–305).

## 2.5 Life Table Transition Probabilities Literature

Various techniques exist to estimate the probability of success in a binary event as a function of one or more predictor variables. One common method involves fitting a logistic regression model and using the model fit to obtain success probability estimates. In the simple version of a logistic regression model in which there is only one continuous predictor variable, $x$, it is assumed that

$$\log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x,$$

where $\pi(x)$ is the probability of success as a function of $x$ (Agresti 2002, p. 122). As Agresti notes, the conversion from $\pi(x)$ to $\log \frac{\pi(x)}{1-\pi(x)}$ is referred to as the logit link, so that a logit model is just another name for a logistic regression model (Agresti 2002, p. 123). The more general representation of the technique is to assume

$$\pi(x) = \Phi(\alpha + \beta x),$$

where $\Phi$ represents a class's standard cdf (Agresti 2002, p. 124). With this more general representation, if $\Phi(c_0) > \Phi(c_1)$ whenever $c_0 > c_1$ (these inequalities are not explicitly defined in a referenced source), the linear model is

$$\Phi^{-1}\left[\pi(x)\right] = \alpha + \beta x$$

(Agresti 2002, p. 124). The logistic regression and probit models are particular cases of this general technique using the standard logistic and standard normal cumulative distribution functions for $\Phi$ (Agresti 2002, pp. 124–125).

In the more general case where the response is multinomial rather than simply binomial, however, the logistic regression approach must be modified. Agresti and Greene explain the baseline-category and cumulative logit models, which are used for models of nominal and ordinal categories of responses, respectively. Agresti and Greene also provide an introduction to other types of multinomial data models (Agresti 2002, chap. 7; Greene 2003, chap. 21).

In the baseline-category logit model, one of the possible multinomial responses is picked to be the baseline category (Agresti 2002, p. 268). As the name suggests, all other responses are compared to this one. Using $J$ to denote the number of possible responses as well as the baseline category, it is assumed that

$$\log \frac{\pi_j(x)}{\pi_J(x)} = \alpha_j + \beta_j x, \ j = 1, \ldots, J-1$$

(Agresti 2002, pp. 267–268; note that the equation has been modified to include only one explanatory variable). Agresti notes that if $\alpha_J$ and $\beta_J$ are set to zero then the above equation would apply for all $j$ (Agresti 2002, p. 271). Consequently,

$$\pi_j(x) = \frac{\exp(\alpha_j + \beta_j x)}{1 + \sum_{k=1}^{J-1} \alpha_k + \beta_k x}$$

for all $j$ (Agresti 2002, p. 268; again, the equation has been modified to include only one explanatory variable).

While the baseline-category logit model is a rather flexible approach to obtain multinomial probability estimates, a different alternative is sought so that the probability of dying can be interpreted in a more straightforward fashion. Instead of modeling a multinomial response in one step, a binomial response could first be modeled, and then a baseline-category logit model could be used. That is, suppose there are $J > 2$ possible responses. In the first step, the model treats a success as

having $j = 1, \ldots, J-1$ so that only the $j = J$ response is considered a failure. Then, letting $\pi_r(x)$ denote the probability of a success,

$$\pi_r(x) = \frac{\exp(\alpha_r + \beta_r x)}{\exp(\alpha_r + \beta_r x) + 1}.$$

In the next step, a conditional baseline-category logit model—conditional on the response not being the $J^{\text{th}}$ response—could be used. This can easily be done by ignoring all occurrences of the $J^{\text{th}}$ response when fitting the baseline-category logit model. Let $\pi_s(x)$ denote the probability of the $s^{\text{th}}$ response, $s = 1, \ldots, J-1$, conditional on the response not being the $J^{\text{th}}$ response. If we set both $\alpha_{J-1}$ and $\beta_{J-1}$ equal to zero so that $s = J-1$ is the baseline category,

$$\pi_s(x) = \frac{\exp(\alpha_s + \beta_s x)}{1 + \sum_{s=1}^{J-2} \exp(\alpha_s + \beta_s x)}.$$

Thus, the probabilities of each response $j = 1, \ldots, J$ would be modeled as follows:

$$\pi_{j=1}(x) = \Pr(\text{response } 1 | \text{ not response } J) \Pr(\text{not response } J) = \pi_{s=1}(x)\pi_r(x)$$

$$\vdots$$

$$\pi_{j=J-1}(x) = \pi_{s=J-1}(x)\pi_r(x)$$

$$\pi_{j=J}(x) = (1 - \pi_r(x)).$$

The estimation of multinomial probabilities is a pertinent issue for the proposed research. In the Hall and Jones (2004b) model, the probability of an individual going from his or her present age and risk-factor group combination (or state) to another specified state (i.e., the transition probability) is assumed to follow a first-order Markov process. The assumption that each transition probability follows a first-order Markov process implies that, conditional on an individual's health level in the present state, the individual's health level in any previous state has no effect on the probability of the individual moving to the specified future state. Because a

person in the state corresponding to risk-factor group "Living" and age $a$ (recall that a state is a unique combination of risk-factor group and age) can go to either the state corresponding to "Living" and age $a + 1$ or the state corresponding to "Dead" in the next period, there is a binary response. In the extended model an individual can go to one of multiple states in the next period, so the response can be considered multinomial. Actual estimation of the probability of going to each of these states as a function of "effective health input" can be done in many ways (Hall and Jones 2004b, p. 13; see also Hall and Jones 2004a, p. 15). The technique recommended for the extended model is to fit a baseline-category logit model to estimate the probabilities of each state in the next period given an individual's state and health level in the current period.

The data necessary to assess whether or not the transition probabilities in the extended model actually follow a first-order Markov process, as will be assumed, could be collected. A way to make the assessment when there are no explanatory variables is described by Bishop, Fienberg, and Holland (1975). While the Markov-process assumption for the Hall and Jones (2004b) model does not necessarily hold, there is even more reason to question the assumption for the extended model. With the original model, if it is assumed that individuals cannot go from a "Dead" state to a "Living" state, all individuals who at time $t$ are in the state corresponding to "Living" and of age $a$ would have gone through the same sequence of states. Yet in the extended model, all people at time $t$ in a given state need not have had the same sequence of states leading up to that point. Therefore, it could make sense to check whether or not knowing an individual's current state captures all the necessary health history information. Nevertheless, the model extension will proceed on the assumption that this characteristic holds. Were the process really a higher-order, albeit finite, Markov chain, the states could be redefined so that the process is a first-order one (as in Ross 2003, pp. 182–183). In that case, the extended model could be

altered.

## 2.6 Optimization Techniques Literature

There are multiple methods of obtaining solutions to problems of optimization. For simple problems, setting up a Lagrangian equation can be very helpful. Nicholson (2002) has a helpful introduction to Lagrangian equations with economic applications in mind. Other methods, such as the use of an exhaustive search or the use of a popular algorithm by Nelder and Mead, are discussed in Haupt and Haupt (2004). For the problem at hand, the function that must be optimized takes a large number of arguments and has dynamic structure (i.e., many time periods are considered, and results from the current time period affect the results for future time periods). The dynamism of the equation as well as the large dimensionality of the function to be optimized suggest that a different optimization approach might be helpful. Hall and Jones (2004b) include an algorithm with their paper which they utilized to obtain numerical answers to optimize the relevant Bellman equation (Hall and Jones 2004b, p. 40; see also Hall and Jones 2004a, p. 40). They were able to use first-order conditions in a relatively simple manner.

This project uses an alternative to such a strong dependence on first-order conditions. Genetic algorithms can identify values for the arguments that provide a close-to-optimal outcome even if certain mathematical properties like continuity are not present (Hamada, Martz, Reese, and Wilson 2001, p. 176). Thus, even if first-order conditions are not sufficient for determining optimizing arguments, a genetic algorithm can be used. Also, because the maximization involved in the proposed model extension is over many parameters, a genetic algorithm approach could be used because it is a relatively time-efficient means of finding a close-to-optimal outcome (C. S. Reese, personal communication, July 12, 2005).

Hamada et al. (2001) provide a very understandable description of genetic al-

gorithms. The basic idea seems to reflect what is thought to go on in evolution—that in mating, chromosomes containing genetic characteristics of the mates stochastically crossover and mutate; individuals with the best-suited chromosomes tend to be more likely to pass their chromosomes along (Hamada et al. 2001; Haupt and Haupt 2004, chap. 1). Genetic algorithm–based optimization considers a pool of proposed sets, where each set is a group of values corresponding to the function's arguments (Hamada et al. 2001, p. 176). Sets from the pool can be stochastically "mated," with the sets being crossed-over and mutated (Hamada et al. 2001, p. 177). Based on genetic algorithm literature, it seems that the algorithm attempts to guide the mating, crossing-over, and mutating processes so that repeated application yields close-to-optimizing argument values.

The implementation of a genetic algorithm requires additional considerations when the objective function has dynamic equality constraints and many arguments over which to optimize. Based on varied readings regarding the genetic algorithm, it seems that for a complex model, such as the extended model proposed in Chapter 3, letting the genetic algorithm produce sets of argument values in an unconstrained fashion will almost certainly create problems by resulting in sets that do not meet the constraints. Adjustments can be made to penalize functions in a variety of manners for not meeting constraints (Sakawa 2002, p. 133), or reference points can be used in conjunction with search points to produce sets in each generation, or iteration, that do meet the constraints, as first done by Michalewicz and Nazhiyath (1995) and improved by Sakawa and Yauchi (1998) (Sakawa and Yauchi 1998, p. 885; also explained thoroughly in Sakawa 2002, chap. 7). While these concepts should be used in obtaining a close-to-optimal solution for the entire extended model, because of the complex nature and abundance of arguments, a simplified version of the extended model that also uses assumed values for unknown parameters will be solved which does not require such techniques and which should be able to be computed much

more quickly. This is described further in Section 3.3.4.

## 2.7 Relation Between Proposed Research and Extant Literature

The proposed research seeks to extend the model presented in Hall and Jones (2004b). Because it is an extension, rather than a replacement, much of their work will be relied upon in making the extension; a similar foundation and many similar elements will be used.

There are additional elements of the proposed work to Hall and Jones' model. Namely, the proposed work will include extra heterogeneity by using more representative agents and can perhaps best be categorized as mimicking the work of Daniel (1993) and others to divide individuals into groups and assume a representative agent in each group. It should be noted that Hall and Jones (2004b) seem to have proceeded in this manner, as they have one representative person for each age group rather than one overall representative. However, the proposed work will incorporate an additional level of heterogeneity by having multiple representative agents for each age group. This allows for there to be multiple risk-factor groups represented in each age. Individuals will be able to move between these risk-factor groups as they age. Also, a quality-adjustment principle will be incorporated when measuring the utility associated with a given health level, as discussed in Section 3.1.3 and Section 3.1.4.

The extended model requires the ability to classify individuals into risk-factor groups so that data which contain information on the age, health spending, and health-related risk factors of individuals can be used in estimating model parameters. Although various techniques exist by which to classify individuals into risk-factor groups, the linear classification technique using prior information seems reasonable. After having assigned individuals into risk-factor groups, successive baseline-category logit models can be used to estimate state-to-state transition probabilities as a function of an individual's health expenditures, as well as of the individual's initial state

and the health technology level. Finally, while perfect optimization will not necessarily be achieved, a genetic algorithm can be used to arrive at a close-to-optimal allocation of resources. However, the example illustrating the use of a genetic algorithm will only attempt to arrive at such a close-to-optimal solution for a very simplified form of the model.

# 3. METHODOLOGY FOR PROPOSED EXTENSION

The work demonstrated in this chapter is (1) the theoretical representation of a possible extension of the Hall and Jones (2004b) model and (2) the implementation of techniques that could be used in some of the parameter estimations and in finding a close-to-optimizing resource allocation; the techniques are implemented on pedagogical examples. The bulk of the work lies in coming up with the theoretical extension of the model, while the examples are included to provide basic examples of classification analysis, successive baseline-category logit model-fittings, and genetic algorithm usage in the context of the extended model. In each instance the examples give results not of primary importance to the extended model due to the nature of the data analyzed or the parameter assumptions.

As for the theoretical representation, many of the Hall and Jones (2004b) model equations are singled out and then modifications to these equations for the additional representative agents case are proposed and justified. In many cases the Hall and Jones (2004b) equations require only minor modifications, but several of them solicit substantial consideration. Thus, coming up with a meaningful modification is nontrivial.

Examples of some of the statistical techniques that could be used in the process of fitting the extended model are described in this chapter, and the results of applying these techniques are mentioned briefly in Chapter 4. It is crucial to make clear that these examples are generally not based on relevant data. Thus, the examples are not intended to represent the required results necessary for fitting the extended model and subsequently finding the best allocation of available resources given the model; instead, the examples are intended to detail several aspects of model fitting and optimization if the appropriate data were accessible. Specifically, data on papers from

*The Federalist* are used to demonstrate classifying units based on certain quantitative characteristics. Baseball plate-appearance data from Dale Murphy's 1987 Major League Baseball (MLB) season is used to demonstrate parameter estimation for successive baseline-category logit models. First, the parameters are estimated for an initial baseline-category logit model. Then the parameters are estimated for another baseline-category logit model, conditional on a certain outcome. Many assumptions, including assumed values for all parameters of the theoretical model extension, are made so that a genetic algorithm can be used in finding close-to-optimizing argument values.

## 3.1 Setup of Extended Model Structure with "Healthy" and "Sick" Risk Groups

We begin by supposing that all individuals of the same age at a given time are not the same with regards to their health as a function of expenditures for health. This is unlike Hall and Jones (2004b) and Hall and Jones (2007), where in each case their model assumes that individuals of the same age are the same. Instead, we suppose that there are two risk-factor groups present in living people, namely "Healthy" and "Sick," so that at any given time those persons who are in the same risk-factor group and have the same age have the same health function, but are allowed to be different from those persons with a different risk-factor group or age at that time. We also allow individuals to move between these two risk groups as time progresses.

Upon dying an individual is no longer considered in the extended model. Because death is the event through which individuals exit the model, a third risk-factor group, "Deceased," is included. The motivation for this third risk-factor group is that it allows individuals to effectively be ignored in all future periods of the model even though the individuals are still in one of the model's states in every future period. An individual's sequence of state memberships over time will be assumed to follow a Markov process; thus, the requirement of a Markov process that individuals always

belong to a state in every future period is satisfied because of this third risk-factor group (see Ross 2003, p. 181).

We use two different notations to denote a health state corresponding to a unique combination of risk-factor group and age (except in the case of "Deceased," for which age is irrelevant). The first notation is to use the subscript $k$ to denote the state; for example, health state $k = 0$ could correspond to "Deceased," while $k = 1$ could correspond to age 0 and in "Healthy" group, $k = 2$ could correspond to age 0 and in "Sick" group, $k = 3$ could correspond to age 1 and in "Healthy" group, $k = 4$ could correspond to age 1 and in "Sick" group, and so on. The second notation, which will often be used, is the pair $(g, a)$ where $g = 1, 2, 3$ represents risk-factor groups "Healthy," "Sick," and "Deceased," respectively, and $a = 0, 1, 2, \ldots$ represents age. Note, however, that when $g = 3$, the age subscript is unnecessary for distinguishing between states because all individuals in the "Deceased" risk-factor group are put in a common state; consequently, the $(g, a)$ representation of the "Deceased" state is $(3, \cdot)$.

It should be noted that there are many possible states in the model. For example, suppose a time period represents one year. The number of possible states depends on whether individuals' attainable ages are assumed to be unbounded or not; if the attainable ages are assumed to be unbounded, there is a countably infinite number of states, while if the attainable ages are assumed to have an upper bound, $A$, then there would be $2(A + 1) + 1$ states (one state for "Deceased" and two other states for each age $0, 1, 2, \ldots, A$). However, three states at most are possible in the next period. Given an individual's current state, say $(g, a)$, the individual's next state will either be $(1, a + 1)$, $(2, a + 1)$, or $(3, \cdot)$.

As previously mentioned, an individual's sequence of health states will be assumed to follow a Markov process, except for those entering states corresponding to age 0, as the model would not have a previous health state for these individuals. For

the distribution of individuals in the states associated with (1) "Healthy" and age 0 and (2) "Sick" and age 0, see the assumption in Section 3.1.5. Because the states are constructed to contain pertinent age and risk-factor group information, it is supposed that the present state adequately captures the health history information needed to calculate the probability of being in any other state in the next time period.

### 3.1.1 Determining Equations for Transition Probabilities and Health Level

A particularly critical assumption in the Hall and Jones (2004b,2007) models is the nature of the deterministic equation which characterizes a person's health level. The structure of the equation is particularly important because the equation itself affects the model in two important ways:

(1) a person's health level in a particular period is one of the arguments in the person's utility function for that period, and

(2) a person's health level is the inverse of the person's probability of dying during that period.

Hall and Jones (2004a,b) use a particular form for the function determining $x_{a,t}$ in their estimation—though not in their theoretical model—and use $1/x_{a,t}$ as the probability that someone of age $a$ at time $t$ is not alive in the next period (Hall and Jones 2004b, pp. 12–13, 18; see also Hall and Jones 2004a, pp. 14–15, 20). There is not a lone, clear-cut extension to the extended model; that is, it is not clear how a one-dimensional health level would determine the probabilities of being in either of two states in the next period; note that only two probabilities are needed to identify all three probabilities because they sum to one. The question then arises as to whether or not the probabilities should be related to the health level. It seems clearly preferable that they be related, especially because in the Hall and Jones (2004b, 2007) model the two properties have such an obvious correspondence.

There are several potential methodologies to formulate the health level equations and the transition probability functions. One methodology is to simultaneously propose health level equations and transition probability functions; this approach seems difficult to implement because of the complexity in simultaneously proposing both types of functions, given that the two types of functions need to have a meaningful relationship to each other. Another methodology is to first propose the health level functions and then to derive the transition probability functions from the health level. A different methodology is to first propose the transition probability functions and then to derive the health level functions from the transition probabilities.

The extended model uses the methodology of first formulating transition probability functions and then deriving health level equations. To justify the use of this approach, it should be noted that Hall and Jones (2004a,b) used mortality data to estimate the parameters in their age-specific health level equation. Indeed, it is not obvious how they would have estimated the parameters without using mortality data. In the extension, there is an additional reason to follow the second proposed course—it is preferable to reduce two probabilities, the third being redundant, to one health level as opposed to expanding one health level into two probabilities, from which the third probability would be deduced.

### 3.1.2 Transition Probabilities Equations

While ideally the defined state-to-state transition probabilities would be a direct analogue of the Hall and Jones (2004a,b) model's defined probabilities, they are not. However, an attempt to make a logical extension has been made. Note that Hall and Jones (2004b, 2007) allow for a general health function in the theoretical model and then use a certain form in the empirical version. The proposed model will differ now in that it will assume a form for the theoretical *and* empirical model. In the empirical model of Hall and Jones (2004a,b), the probability of an individual of age $a$ in time

period $t$ dying before the next time period is given by

$$\frac{1}{x_{a,t}} = \frac{1}{A_a(z_t h_{a,t})^{\theta_a}},$$

(from combining equations in Hall and Jones 2004b, pp. 12–13, 18; see also Hall and Jones 2004a, pp. 14–15, 20). Upon further investigation of the Hall and Jones (2004a,b) model, it seems problematic that if $x_{a,t}$ is less than one, individuals with that health level would have a probability greater than one of dying before the next period. Noting estimates provided by Hall and Jones (2004b, 2007), sufficiently small but positive levels of effective health expenditures would imply health levels less than one and thus probabilities larger than one of not living to the next period. However, had Hall and Jones (2004a,b) modeled the probability of dying in the present period as

$$\frac{1}{A_a(z_t h_{a,t})^{\theta_a} + 1},$$

then positive values of $A_a$ and $\theta_a$ would not allow any positive level of health inputs to give a value in the denominator less than one, though the denominator would not be bounded above. However, because of constraints on the values $h_{a,t}$ can take on in any finite time period, the denominator is effectively bounded above. This means the probability of dying in the present period is prohibited from being larger than one and less than zero.

Modeling the rate of dying for an age group in this fashion has another advantage besides excluding probabilities less than zero or greater than one. The probability in the preceding paragraph can be expressed as a logit model rather easily. To see this, first note that

$$A_a(z_t h_{a,t})^{\theta_a} = \exp(\ln(A_a) + \theta_a \ln(z_t h_{a,t}))$$

(compare with nearly equivalent equation on Hall and Jones 2004b, p. 19; see also Hall and Jones 2004a, p. 21); thus, letting $p_{a,t}(z_t h_{a,t})$ denote the probability that an

individual is alive in the next period,

$$\ln\left(\frac{p_{a,t}(z_t h_{a,t})}{1-p_{a,t}(z_t h_{a,t})}\right) = \ln(A_a) + \theta_a \ln(z_t h_{a,t}).$$

Note that this logit model is in the form $\alpha + \beta x$, where $\alpha = \ln(A_a)$, $\beta = \theta_a$, and $x = \ln(z_t h_{a,t})$.

For the extended theoretical—and, equivalently, empirical—model probabilities, recall that risk groups $g = 1, 2, 3$ correspond to the "Healthy," "Sick," and "Deceased" groups. Let $P_{gg',a,t}(z_t h_{g,a,t})$ denote the probability that an individual of age $a$ with effective health expenditures $z_t h_{g,a,t}$ and in risk-factor group $g$ in time period $t$ is in risk-factor group $g'$ in the next period with age $a+1$ and time $t+1$. Then the probability that an individual will be alive in the next period (that is, in risk-factor group 1 or 2 in the next period) is denoted by $P_{g(live),a,t}(z_t h_{g,a,t}) = P_{g1,a,t} + P_{g2,a,t}$. The extended model assumes that the probabilities $P_{g1,a,t}$ and $P_{g2,a,t}$ are those which would be obtained by first using a logit model to model the probability of a person living to the next period, and then using a baseline-category logit model to estimate the conditional probabilities of being in each health state given that the person is alive in the next period. Explicitly, the model assumes that the probability of living to the next period follows the relation

$$P_{g(live),a,t}(z_t h_{g,a,t}) = \frac{\exp(\alpha_{g(live),a} + \beta_{g(live),a} \ln(z_t h_{g,a,t}))}{1 + \exp(\alpha_{g(live),a} + \beta_{g(live),a} \ln(z_t h_{g,a,t}))}.$$

Letting $P_{g(g'|live),a,t}(z_t h_{g,a,t})$ represent the probability of being in risk-factor group 1 in the next period conditional on being alive (i.e., not in risk group 3) in the next period, the extended model assumes that this probability follows the relation

$$P_{g(g'|live),a,t}(z_t h_{g,a,t}) = \frac{\exp(\alpha_{g(g'|live),a} + \beta_{g(g'|live),a} \ln(z_t h_{g,a,t}))}{\sum_{i=1}^{2} \exp(\alpha_{g(i|live),a} + \beta_{g(i|live),a} \ln(z_t h_{g,a,t}))}.$$

The extended model sets $\alpha_{g(2|live),a}$ and $\beta_{g(2|live),a}$ equal to zero for all $(g, a)$ combinations so that the baseline category always has "Sick" as the risk group. Such a constraint is made for purposes of identifiability (see Agresti 2002, p. 271). The state

that a person will belong to in the next time period, given that the person will be alive in the next time period, can be considered as a binomial response because there are only two possible states; thus, the last equation corresponds to a probability from a logit model. The conditional probabilities in the last equation are depicted as corresponding to probabilities from a baseline-category logit model, rather than simply a logit model, because the baseline-category logit model easily admits an expansion of the model to contain more risk-factor groups.

Therefore, the state-to-state transition probabilities are assumed to be as follows:

$$P_{g1,a,t}(z_t h_{g,a,t}) = P_{g(live),a,t}(z_t h_{g,a,t}) P_{g(1|live),a,t}(z_t h_{g,a,t}),$$

$$P_{g2,a,t}(z_t h_{g,a,t}) = P_{g(live),a,t}(z_t h_{g,a,t}) P_{g(2|live),a,t}(z_t h_{g,a,t}), \text{ and}$$

$$P_{g3,a,t}(z_t h_{g,a,t}) = (1 - P_{g(live),a,t}(z_t h_{g,a,t}))$$

for $g = 1, 2$. Recall that upon dying individuals are no longer considered in the model. The convention used to restrict individuals who have died from further impacting the model is to admit them to an absorbing state (that is, a state that is never left once it is entered). This absorbing state is ignored in the model, thereby preventing individuals who have died from affecting the model. This absorbing state is referred to as the "Deceased" state because death is the event through which individuals enter this state. Because individuals in the "Deceased" state remain in the "Deceased" state, and because the "Deceased" state corresponds to risk-factor group 3, $P_{33} = 1$ by construct, regardless of age or time.

Note that many state-to-state combinations are not accounted for in the equations (such as going from the state corresponding to "Sick" and age 2 to the state corresponding to "Healthy" and age 2) because such transitions are impossible, and thus are known to have zero probability without the need for an equation to relay this information.

It is assumed that the level of health technology, $z_t$, is not health-state specific. If an accurate measure of the level of health technology available to each health state in each period were identified, this assumption could easily be relaxed.

### 3.1.3 Health Level Equation

Now that a relation for determining the state-to-state transition probabilities is in place, we turn our attention to relating the health level to these probabilities. The sole purpose of doing so is to have a measure which can be included in the utility function so that health does not merely impact life's length without consideration of life's enjoyability. For instance, suppose two individuals both live to age 70 and have identical consumption in each time period. If the first individual was in the "Healthy" group at each age until the time of death, but the second individual spent 25 years in the "Sick" group before dying, clearly the first individual should have a greater amount of total utility over those 70 years than the second individual.

While the motivation for using the health level in the utility function is apparent, a decision must be made about how to link the health level to the transition probabilities. Though there may not be one correct way to define the health level, it seems natural to use a principle related to using the expected number of quality-adjusted life years an individual has remaining rather than just the expected number of additional years a person will live.

Although Hall and Jones (2004b, 2007) probably did not intend this interpretation, the health level's impact on utility in their models can be characterized as the psychological effect on an individual's well-being that is due to the individual's survival probability. For example, an individual might have peace of mind by knowing that the probability of living until the next time period is high. This view prompts the assumption that the following relation defines an individual's health level as it

pertains to utility:

$$x_{g,a,t}(z_t h_{g,a,t}) = \frac{1}{1 - P_{g1,a,t}(z_t h_{g,a,t}) - \delta P_{g2,a,t}(z_t h_{g,a,t})}.$$

Here the parameter $\delta$ represents how much an individual values an increase in the probability of being in a "Sick" state if the probability of being in a "Healthy" state is kept constant relative to how much that individual values an increase in the probability of being in a "Healthy" state if the probability of being in a "Sick" state is held constant. It is assumed that $0 \leq \delta \leq 1$. In order to better understand this equation, consider the implication of various $\delta$ values. If $\delta$ is equal to zero, an individual is indifferent between dying and being in the "Sick" risk-factor group, in which case that individual's health level is determined by the probability of being in the "Healthy" group in the next period. If $\delta$ is equal to one, an individual has no preference for being "Sick" versus being "Healthy," in which case the individual's health level is determined by the probability of being alive in the next period, as in the Hall and Jones (2004b, 2007) health level equation. If $0 < \delta < 1$, an individual prefers being in the "Healthy" group to being in the "Sick" group but prefers being in the "Sick" group to dying.

### 3.1.4 Utility Equation

Because individuals are assumed to be exactly the same only if they are in the same health state at the same time, the utility function could be state- and time-dependent. Aside from its dependence on state and time instead of age and time, the utility function has the same representation as a function of consumption spending and health spending as it does in the Hall and Jones (2004a,b) model. Another notable change is that the health level is calculated in a manner different from the Hall and Jones (2004b, 2007) model. The utility equation is

$$u_{k,t}(c_{k,t}, x_{k,t}) = b_{k,t} + \frac{c_{k,t}^{1-\gamma}}{1-\gamma} + \alpha \frac{x_{k,t}^{1-\sigma}}{1-\sigma}$$

(nearly identical to Hall and Jones 2004b, p. 13; see also Hall and Jones 2004a, p. 15). There are three components of the utility function. The first part, $b_{k,t}$, represents the base amount of utility associated with being in health state $k$ in time period $t$. The second part, $\frac{c_{k,t}^{1-\gamma}}{1-\gamma}$, represents the effect due to consumption spending in time $t$. The third part, $\alpha\frac{x_{k,t}^{1-\sigma}}{1-\sigma}$, represents the effect due to the health level in time period $t$ (Hall and Jones 2004b, p. 13; Hall and Jones 2004a, p. 15). The third part can be weighted more or less by varying the value of $\alpha$ (Hall and Jones 2004b, p. 27; Hall and Jones 2004a, p. 29).

### 3.1.5 Objective Function and Associated Constraints

Suppose that given $N_t$—that is, a vector containing the number of people in each health state at time $t$—the decision-maker seeks to find $c_{k,t}$ and $h_{k,t}$ for all $k$ not equal to zero so as to maximize the following Bellman equation:

$$V_t(N_t) = \max_{h_{k,t},c_{k,t}} \left( \sum_{k=1}^{\infty} N_{k,t} u_{k,t}(c_{k,t}, x_{k,t}) + \beta V_{t+1}(N_{t+1}) \right)$$

(nearly identical to Hall and Jones 2004b, p. 14; see also Hall and Jones, 2004a, p. 16). Note that the summation does not include $k = 0$ because this value corresponds to those in the "Deceased" state, and it is assumed that individuals in this state are ignored.

The imposed constraints, meant to be analogous to those of Hall and Jones (2004a,b), are now included. Note first that one of the constraints is given by the health function (Hall and Jones 2004b, p. 15; Hall and Jones 2004a, p. 17), which function for our extended model is given in Section 3.1.3 and which depends itself on equations in Section 3.1.2.

The constraint regarding the total output in any given period is

$$\sum_{k=1}^{\infty} N_{k,t}(y_t - c_{k,t} - h_{k,t}) = 0$$

(nearly identical to Hall and Jones 2004b, pp. 14–15; Hall and Jones 2004a, pp. 16–17). This constraint implies that the total amount of output in a given time period is exactly equal to the sum of the total amount of income spent on consumption and the total amount of income spent on health. There could be a net surplus of output in some states, but this would mean a net shortage in at least one other state. As in Hall and Jones (2004b, 2007), it will be assumed that every living person produces the same amount of output in a given time period, or that $y_t$ does not depend on the health state, provided that the state is not "Deceased."

The constraint on how the per capita income changes over time is

$$y_{t+1} = e^{g_y} y_t$$

(Hall and Jones 2004b, pp. 14–15; Hall and Jones 2007, p. 50). The amount of income any person, irrespective of age and health group type, earns in time $t + 1$ is some constant times the amount of income a person earned in time $t$. This implies per capita income has exponential growth, with the growth rate being denoted by $g_y$ (Hall and Jones 2004b, p. 15; Hall and Jones 2007, pp. 50–51).

The constraint on how the health technology level changes over time is

$$z_{t+1} = e^{g_z} z_t$$

(Hall and Jones 2004b, pp. 14–15; Hall and Jones 2004a, p. 17). The amount of health technology available in time $t + 1$ is the product of the amount of health technology available in time $t$ and a constant. This implies that the health technology level has exponential growth with growth rate $g_z$ (Hall and Jones 2004b, pp. 15, 18; Hall and Jones 2004a, pp. 17, 20).

And of notable distinction, there are relations to determine the number of people of each age and risk-factor group combination in the next period given information on the present period:

$$N_{g,a+1,t+1} = \sum_{g'=1}^{3} N_{g',a,t} P_{g'g,a,t} (z_t h_{g',a,t})$$

$$N_{g,0,t} = N_{g,0}$$

(similar to Hall and Jones 2004b, p. 14; Hall and Jones 2007, p. 50). The first of these relations specifies that the number of people in the health state corresponding to risk-factor group $g$ and age $a+1$ at time $t+1$ is equal to the sum over each risk-factor group $g'$ of the number of people in risk-factor group $g'$ and of age $a$ at time $t$ multiplied by the probability that such a person will be in risk-factor group $g$ in the next period. As previously mentioned, the extended model will impose the probability of going from the "Deceased" group to the "Healthy" or "Sick" risk-factor groups as being zero. The second constraint states that the same number of people are born into each risk-factor group each year (Hall and Jones 2004b, p. 15; Hall and Jones 2007, p. 50).

## 3.2 Note on Extension to More Risk-Factor Groups

It is relatively straightforward to extend the theoretical model to include a larger, albeit finite, number of risk groups. Let $G$ indicate the number of risk-factor groups, including the "Deceased" risk-factor group, to be included. Assume that risk-factor group 1 is the healthiest risk group and risk group $G$ is the "Deceased" group. The transition probabilities would be given by

$$P_{g1,a,t}(z_t h_{g,a,t}) = P_{g(live),a,t}(z_t h_{g,a,t}) P_{g(1|live),a,t}(z_t h_{g,a,t})$$

$$\vdots$$

$$P_{g(G-1),a,t}(z_t h_{g,a,t}) = P_{g(live),a,t}(z_t h_{g,a,t}) P_{g((G-1)|live),a,t}(z_t h_{g,a,t})$$

$$P_{gG,a,t}(z_t h_{g,a,t}) = (1 - P_{g(live),a,t}(z_t h_{g,a,t})),$$

where

$$P_{g(live),a,t}(z_t h_{g,a,t}) = \frac{\exp \alpha_{g(live),a} + \beta_{g(live),a} \ln(z_t h_{g,a,t})}{1 + \exp \alpha_{g(live),a} + \beta_{g(live),a} \ln(z_t h_{g,a,t})}$$

and

$$P_{g(g'|live),a,t}(z_t h_{g,a,t}) = \frac{\exp \alpha_{g(g'|live),a} + \beta_{g(g'|live),a} \ln(z_t h_{g,a,t})}{\sum_{i=1}^{G-1} \exp \alpha_{g(i|live),a} + \beta_{g(i|live),a} \ln(z_t h_{g,a,t})}$$

for $g = 1, \ldots, G-1$. Because it is assumed that individuals in the "Deceased" state remain in the "Deceased" state, $P_{GG} = 1$ by construct. We could set $\alpha_{g((G-1)|live),a}$ and $\beta_{g((G-1)|live),a}$ equal to zero so that the baseline category for the risk groups associated with being alive is risk group $G-1$.

Changing the health level equation for use in the utility function is not quite as straightforward, but could be done by using the following form:

$$x_{g,a,t}(z_t h_{g,a,t}) = \frac{1}{1 - P_{g1,a,t}(z_t h_{g,a,t}) - \sum_{g'=2}^{G-1} \delta_{g'} P_{gg',a,t}(z_t h_g, a, t)}.$$

Here, it would be assumed that $0 \leq \delta_i \leq 1$ for $i = 2, \ldots, G-1$ so that the probability of being in any of these risk-factor groups needs to be able to be given a weight relative to the first risk group, which is assumed to be the healthiest.

By letting $k$ correspond to a health state, or unique $(g, a)$ combination, with all "Deceased" risk group members being put in state $k = 0$, many of the equations already mentioned need no further modification for this more general case; the only equation requiring modification is that used in determining the number of people in each state for period $t+1$, which uses $G$ instead of 3 in the sum.

$$u_{k,t}(c_{k,t}, x_{,t}) = b_{k,t} + \frac{c_{k,t}^{1-\gamma}}{1-\gamma} + \alpha \frac{x_{k,t}^{1-\sigma}}{1-\sigma}$$

$$V_t(N_t) = \max_{h_{k,t}, c_{k,t}} \left( \sum_{k=1}^{\infty} N_{k,t} u_{k,t}(c_{k,t}, x_{k,t}) + \beta V_{t+1}(N_{t+1}) \right)$$

$$\sum_{k=1}^{\infty} N_{k,t}(y_t - c_{k,t} - h_{k,t}) = 0,$$

$$y_{t+1} = e^{g_y} y_t$$

$$z_{t+1} = e^{g_z} z_t$$

$$N_{g,0,t} = N_{g,0}$$

$$N_{g,a+1,t+1} = \sum_{g'=1}^{G} N_{g',a,t} P_{g'g,a,t}(z_t h_{g',a,t})$$

Again, these equations are either identical to—in the case of the equations for $y_{t+1}$ and for $z_{t+1}$—or based on equations in Hall and Jones (2004b). Also, note that the last two equations would only need to be applied for $g = 1, \ldots, G-1$, because individuals in the "Deceased" risk-factor group would no longer affect the optimization decision as it is set up.

In the remainder of the document the extension with risk groups "Healthy," "Sick," and "Deceased" will be implied when referring to the extended model.

## 3.3 Obtaining Values for Unknown Parameters

Essential equations of the extended model, including the formulas for determining the probabilities as a function of effective health expenditures, have been symbolically depicted. This has the effect of demonstrating how the Hall and Jones (2004b) theoretical model could be adjusted to have more agent heterogeneity by using more health states with a representative agent in each one and allowing for more options as far as future permissible health states. However, note that a specific form for the health level equations has been imposed in the theoretical version of the extended model, whereas Hall and Jones (2004b, 2007) wait to use specific forms until their empirical models in the latter portion of their papers.

One of the first data-driven tasks needing to be performed in applying the model to the U.S. health economy is estimating the transition probability parameters. The estimates could then be used in place of the parameters in subsequent calculations. This task requires two steps: (1) finding a way to appropriately classify individuals into health states, and (2) using those states to estimate the transition probability parameters; the transition probability parameter estimates result from fitting a logit model as well as a baseline-category logit model. Because a logit model is a special

case of a baseline-category logit model, this procedure fits two baseline-category logit models.

### 3.3.1 Classifying Individuals into States

The ability to fit baseline-category logit models to estimate state-to-state transition probability parameters depends on having the health state in a given time period, the effective health expenditures for that time period, and the health state in the next time period for each person used in the models. A person's health state is not always obvious to measure directly. The difficulty lies not in determining age at a given time, but in determining the risk-factor group at that time.

The means to classify the health state based on straightforward measurements is necessary, but coming up with those means is restricted by the lack of a reference set from each state. Without a reference set from each state that can be studied, a classification rule cannot be formed. In order to obtain a reference set, the following approach could, but will not, be used: randomly select a sufficiently large number of people of each age. Take a number of health-related measurements on each of them (e.g., cholesterol level, presence or absence of malaria, etc.). Have a panel of several experienced physicians diagnose each of these people as either "Healthy" or "Sick" after reviewing these measurements. The majority vote would decide what that person's "true" risk-factor group at the time is considered to be. Combining this diagnosis with the age would identify the state the person is considered to be in at that time.

The illustration of how the data could be analyzed is described; the analysis is contingent on having reference data containing health measurements and health states. Suppose that all of the random variables in an individual's response vector of health measurements are continuous or categorical with a finite number of categories. Note that such categorical variables could be recoded using dummy variables for

nominal categories and using the ranked value of the category for ordinal categories; the recoded variables can be analyzed as continuous variables (Rencher 2002, p. 315).

The above rule is now illustrated on a data set of no particular meaning for this problem other than as a vehicle for showing how the technique is used. Information on the sources of this data, as well as the sources of all other data used in this paper, can be found in Appendix A. The data set used to show the classification technique was originally used in Collins, Kaufer, Vlachos, Butler, and Ishizaki (2004). It contains quantitative information on various characteristics of the writing in different papers from *The Federalist* as well as the author of each text, whether it be Madison, Hamilton, or "Disputed" (`http://lib.stat.cmu.edu/datasets/federalistpapers.txt`; Collins et al., 2004).

As for the cause of twelve of these papers having a "Disputed" author, Rencher (2002) explains that there are twelve papers that both of the aforementioned authors claimed to have written (Rencher 2002, p. 300). For the present purposes, this information is ignored.

Suppose all of these papers were written by either Madison or Hamilton and that the texts originally had the author's name on them but a random sample of the papers had the name of the author removed, so the author is grouped under "Disputed." An attempt is made to classify the author of each "Disputed" paper; it is worth mentioning that such classification has been a focus of research by Mosteller and Wallace (1984), as cited in Rencher (2002, p. 300), and Collins et al. (2004).

The "Disputed" papers have already been assumed to be randomly chosen from all of the papers; a further assumption made is that the proportion of papers with a known author that were written by Madison represents the proportion of *all* of the papers that were written by Madison. This assumption implies that without knowing anything about a particular paper, the probability of it having been written by a given author is exactly equal to the proportion of papers written by that author in

the collection of all papers for which the author identification is explicit.

We also assume the distribution of the vector representing all of the characteristics to be measured on a paper from a given author has a multivariate normal distribution for both authors, and that the covariance matrices for these random vectors are exactly the same. We will then use the linear classification criterion as stated in Section 2.4.2 to classify each of the "Disputed" papers as having been written by either Madison or Hamilton. This will be done using all available measurements. The *DISCRIM Procedure* in SAS 9.1 will be used to make the classifications; the code is included in Appendix B.1.

This example illustrates the classification technique necessary in the extended version of the Hall and Jones (2004b) model because it deals with classification into one of two candidate populations where it is assumed that the measured variables from each population follow a multivariate normal distribution with no difference in the covariance matrix of either population. Also, the probability of being in a particular candidate population prior to seeing the realization of the random variables in a response vector has been assumed to be represented by the proportion of members of that population in the reference-set sample. Similarly, we assume in the extended model that the age-specific sample proportions of individuals in each risk-factor group equal the age-specific *a priori* probabilities of each health state.

### 3.3.2 Estimating Baseline-Category Logit Model Parameters

Fitting a baseline-category logit model with four possibilities for the state in the next period will now be discussed using baseball data. A compilation of data sets contains information for every plate appearance in a season for many Major League Baseball (MLB) players in the 1987-1990 MLB seasons. Two important variables in the data set are (1) the season Earned Run Average (ERA) of the pitcher who is pitching in that plate appearance and (2) the discrete outcome of that plate

appearance, with the categories for the outcome being a non-sacrifice out, a one-base hit, a multiple-base hit, or either a sacrifice or a walk (from file BB_Readme by Albright, accessible through `http://www.dartmouth.edu/~chance/teaching_aids/data/baseball.zip`).

A baseline-category logit model with only an intercept and slope coefficient for ERA will be fitted for Dale Murphy's 1987 plate appearances. While the code could be used in fitting data from other players and for multiple seasons, it will suffice to use the data from one player for one season. The first baseline-category logit model will be used to estimate the probability of not getting a non-sacrifice out (analogous to not dying in the extended model). The baseline category for this first model will be a non-sacrifice out and the other categories will be grouped together as "non-outs." The next baseline-category logit model will include only those observations which did not result in Murphy getting a non-sacrifice out and will use getting either a sacrifice or a walk as the baseline category. The code is included in Appendix B.1 and was run in SAS 9.1 using the *Logistic Procedure*.

This example shows how a baseline-category logit model with more than two possible responses can be fitted when there is a continuous explanatory variable. In the extended model, it is assumed that individuals in a particular health state at a particular time have their probability of being in a given health state in the next period determined by their effective health expenditures in that period, but the specific function is state dependent. Thus, if the necessary data were collected, all the observations in which an individual was in a given health state in one period could be used to estimate the parameters of the transition probabilities specific to that initial state as a function of effective health expenditures. Two different baseline-category logit models could be fitted for each health state—the first model for the probability of living and the second model for the conditional probability of each risk group (given that the next health state is not "Deceased"). The estimated parameters would be

assumed to be the true values, thus defining what each transition probability is. A suggested assumption for determining the health technology level for each period is to assume like Hall and Jones (2004a,b) that $g_z = 0.01$, so that $z_t = z_0 e^{(0.01)t}$. And $z_0$ could be set such that the health level in 2005 is one, so that $z_t = e^{(0.01)(t-2005)}$. Doing this would allow health expenditure amounts, which would need to be collected, to be converted into effective health expenditure amounts so that the models could be fitted.

### 3.3.3 Obtaining Values for Other Parameters

After estimating the parameters of the state-to-state transition probability matrix and assuming they are the true values, there are still quantities that are yet to be determined. These include $\sigma$, $\alpha$, $\gamma$, and $b_{k,t}$ (all components of the utility function); $\beta$, the parameter by which utility in the next period is discounted relative to the present period; and $\delta$, the parameter reflecting the desirability of transferring probability from being in the "Deceased" risk-factor group to the "Sick" risk-factor group in the next period relative to transferring probability from being in the "Deceased" risk-factor group to the "Healthy" risk-factor group in the next period. While none of these parameters will be estimated for this project, some of these parameters have been estimated by Hall and Jones (2004b), albeit in a different context because of the model differences.

Though details of a possible method for estimating each of these parameters will not be explained, several possibilities exist for how this could, but will not, be done. For instance, relevant estimates by Hall and Jones (2004b) could be used where appropriate. One approach, at least for the $b_{k,t}$ parameters, would be to remove the time-period specificity (i.e., use only $b_k$) and to use psychometric techniques—for example, asking a random sampling of individuals carefully-designed questions which reveal their preferences in a multitude of different economic and health-related

situations—from which these parameters might be estimated. Yet, as previously stated, no attempt will be made to estimate these parameters in the extended model for this project. Also, it is noteworthy that Hall and Jones replaced $b_{a,t}$ with $b$ in the later version of their model (Hall and Jones 2007, p. 49).

We now briefly discuss four other parameters: $g_y$, $y_0$, $N_{1,0}$, and $N_{2,0}$. Following Hall and Jones (2004b), we will assume that the real per-capita income in the macroeconomy will grow at 2.31% annually in the future (Hall and Jones 2004b, p. 33; Hall and Jones 2007, p. 64). Hence, log(1.0231), or 0.02284, will be the assumed value for $g_y$. The value for $y_0$ would be the per-capita income in time period $t = 0$. For convenience in implementing the genetic algorithm in a simplified setting, we take the value of $y_0$ to be equal to one—this implies that $h_{g,a,t}$ and $c_{g,a,t}$ would need to be expressed throughout all time periods as real health and consumption expenditures by people in risk group $g$ and of age $a$ in period $t$ divided by the real 2005 per-capita income, since $t = 0$ will denote the five-year time period beginning in 2005. Also, $N_{1,0}$ and $N_{2,0}$, the number of people born into each risk-factor group per year, could be assumed to take on values chosen based on the proportion of infants classified into each of the risk-factor groups (if the procedure in Section 3.3.1 were to be undertaken) and the average number of people born during the last five or ten years. Hall and Jones (2004a,b) accounted for neither immigration nor emigration in the economy (Hall and Jones 2004b, p. 34; see also Hall and Jones 2004a, p. 35). Similarly, these factors will not be accounted for in Section 3.3.4.

### 3.3.4 Using Genetic Algorithm in Simplified Problem

The genetic algorithm in this instance could be very involved, as this is a very complex problem with a large number of parameters and dynamic constraints. Therefore, for the purpose of showing the genetic algorithm, the problem is simplified in that assumed values are assigned to all unknown parameters. Further simplifications

of the problem result by making some additional assumptions and restrictions. For example, ages and time are grouped into five-year blocks (Hall and Jones 2004b, 2007). This restriction induces several changes, including an important distinction for notation. The genetic algorithm will be used to obtain a close-to-optimal solution beginning in the year 2005, but this will hereafter correspond to $t = 0$. Also, the period $t = 1$ corresponds to the year 2010, not 2006, as time periods are five years long. This in turn means that the age groups will be grouped into five-year blocks (i.e., 0-4, 5-9, etc). The genetic algorithm will be run until year 2100 of the model (or time period 19). The $b_{k,t}$ values are assumed to be time invariant; likewise, Hall and Jones (2004,b) used future values of $b_{a,t}$ that did not depend on time (Hall and Jones 2004b, pp. 30–32; Hall and Jones 2004a, pp. 32–34).

A further assumption that will be made is that all individuals in a given risk group who are age 115 or greater are in the same health state. This assumption does not seem overly restrictive, as not many individuals would be expected to live this long. However, it is made so as to have a finite number of arguments over which to optimize without requiring that all individuals die before a certain age.

Another important assumption, which is made in this example so that the problem is easier to solve, is that $\beta$ equals zero. While this assumption would be undesirable for the theoretical model, it is made in this example of fitting the empirical model because this assumption simplifies the task of demonstrating the use of a genetic algorithm. This assumption implies that the optimization in each period is only over the total amount of utility in that period and does not include some proportion of the total amount of utility in any future periods. This allows the optimization problem to be considered one period at a time. After obtaining argument values in one time period, the argument values are incorporated in the model to set the stage for the next period's optimization problem.

Keeping in mind that the corresponding results should not be interpreted—the

illustration of the technique is what is important—Table 3.1, Table 3.2, and Table 3.3 represent the values for unknown parameters which will be assumed in demonstrating the genetic algorithm.

The values assumed for each of the parameters are not claimed to be true values for any real economy. The selection of assumed values is briefly explained. The reason for setting $\beta$ equal to zero is to reduce the number of arguments simultaneously optimized. The reason for setting $\delta$ equal to 0.7 is so that being sick is substantially worse than being healthy, but being sick is very preferable to being dead. The values of $\alpha$, $\sigma$, and $\gamma$ are picked to be close to values Hall and Jones (2004b) assumed in various circumstances. The recursive relations for $z_t$ and $y_t$ were discussed in Section 3.3.2 and Section 3.3.3, respectively. The values for the remaining parameters were sometimes influenced by values used by Hall and Jones (2004b) and at other times were chosen arbitrarily.

The genetic algorithm to be used based on these parameters is now described. As noted in Section 2.6, the genetic algorithm mimics biological processes of mating, crossing-over, and mutation. While not all details are included here, the following algorithm summarizes the steps taken.

```
For time period i (i=0, 1, 2, ..., 19) {
  Calculate available income for time period
  For run j (j=1, 2, 3, 4, 5) {
    1.  Create 1000 initial sets of argument values using
        (pseudo)random numbers subsequently scaled to satisfy
        the constraints (up to round-off error).
    2.  Evaluate implied overall utility level to permit
        ranking of sets.
    3.  Select sets for crossover based on ranking and switch
        set values at randomly selected locations to generate
        1000 more sets.
    4.  Mutate sets from step 2 by scaling randomly selected
        values by random values to generate 1000 more sets.
    5.  Rescale all argument values to make sure income
        constraint is met for each of the 3000 sets and
        evaluate implied overall utility level for all 3000
```

Table 3.1: Assumptions for unknown probability parameters so genetic algorithm can be demonstrated

| Parameter | Description | Assumed value in genetic algorithm demonstration |
|---|---|---|
| $\alpha_{1(live),a}$ | Age-group-specific intercept in logit model for probability of being alive in next period (five years later) if in healthy group now | (5.2, 4.9, 4.6, . . ., -1.4, -1.7) for age groups (0–4, 5–9, 10–14, . . ., 110–114, 115+) |
| $\alpha_{2(live),a}$ | Age-group-specific intercept in logit model for probability of being alive in next period if in sick group now | (4.2, 3.9, 3.6, . . ., -2.4, -2.7) for age groups (0–4, 5–9, 10–14, . . ., 110–114, 115+) |
| $\beta_{1(live),a}$ | Age-group-specific slope for "effective health spending amount" in logit model for probability of being alive in next period if in healthy group now | (0.48, 0.46, 0.44, . . ., 0.04, 0.02) for age groups (0–4, 5–9, 10–14, . . ., 110–114, 115+) |
| $\beta_{2(live),a}$ | Age-group-specific slope in logit model for probability of being alive in next period if in sick group now | (0.48, 0.46, 0.44, . . ., 0.04, 0.02) for age groups (0–4, 5–9, 10–14, . . ., 110–114, 115+) |
| $\alpha_{1(1\|live),a}$ | Age-group-specific intercept in logit model for probability of being in healthy group in next period if in healthy group now, conditional on being alive in next period | (5.2, 4.9, 4.6, . . ., -1.4, -1.7) for age groups (0–4, 5–9, 10–14, . . ., 110–114, 115+) |
| $\alpha_{2(1\|live),a}$ | Age-group-specific intercept in logit model for probability of being in healthy group in next period if in sick group now, conditional on being alive in next period | (4.2, 3.9, 3.6, . . ., -2.4, -2.7) for age groups (0–4, 5–9, 10–14, . . ., 110–114, 115+) |
| $\beta_{1(1\|live),a}$ | Age-group-specific slope in logit model for probability of being in healthy group in next period if in healthy group now, conditional on being alive in next period | (0.48, 0.46, 0.44, . . ., 0.04, 0.02) for age groups (0–4, 5–9, 10–14, . . ., 110–114, 115+) |
| $\beta_{2(1\|live),a}$ | Age-group-specific slope in logit model for probability of being alive in next period if in sick group now, conditional on being alive in next period | (0.48, 0.46, 0.44, . . ., 0.04, 0.02) for age groups (0–4, 5–9, 10–14, . . ., 110–114, 115+) |

Table 3.2: Assumptions for unknown utility, health level, and discount parameters so genetic algorithm can be demonstrated

| Parameter | Description | Assumed value in genetic algorithm demonstration |
|---|---|---|
| $b_{g,a,t}$ (same as $b_{k,t}$ for some $k$) | Parameter in utility function associated with risk-factor group $g$ and age $a$ for time period $t$ (but will make no difference on optimization decision if $\beta = 0$ | 15 for all ages and risk-factor groups (except "Deceased" group) |
| $\alpha$ | Parameter in utility function for health level | 2 |
| $\sigma$ | Parameter in utility function for health level | 1.7 |
| $\gamma$ | Parameter in utility function for consumption amount | 1.6 |
| $\delta$ | Parameter in health level equation | 0.7 |
| $\beta$ | Discount parameter | 0 |

Table 3.3: Assumptions for other unknown model parameters so genetic algorithm can be demonstrated

| Parameter | Description | Assumed value in genetic algorithm demonstration |
|---|---|---|
| $y_t$ | Standardized real per-capita income for (five-year) period $t$ in units of 2005 real per capita income | $\exp(0.02284 \times 5t)$ because yearly growth rate assumed to be 2.31%, and five years in a period |
| $z_t$ | Standardized health technology level for period $t$ in units of 2005 health technology amount | $\exp(0.01 \times 5t)$ because yearly growth rate assumed to be 1%, and five years in a period |
| $N_{1,0}$ | Number of people in risk group 1 ("Healthy") and in age group 0–4 every period | 10,000,000 |
| $N_{2,0}$ | Number of people in risk group 2 ("Sick") and in age group 0–4 every period | 1,000,000 |
| $N_{1,a,0}$ | Number of people in "Healthy" risk group and in age group $a$ in period 0 | In millions: (10.00, 9.95, 9.90, 9.85, 9.80, 9.75, 9.70, 9.65, 9.60, 9.55, 9.45, 9.10, 8.50, 7.50, 6.00, 2.00, 1.00, 0.20, 0.12, 0.08, 0.01, 0.001, 0, 0) for age groups (0–4, 5–9, 10–14, ..., 110–114, 115+) |
| $N_{2,a,0}$ | Number of people in "Sick" risk group and in age group $a$ in period 0 | In millions: (1.00, 1.00, 1.00, 1.00, 1.00, 1.30, 1.50, 1.60, 1.70, 1.80, 2.00, 2.05, 2.00, 2.50, 2.50, 2.00, 0.50, 0.50, 0.50, 0.30, 0.10, 0.002, 0.0001, 0) for age groups (0–4, 5–9, 10–14, ..., 110–114, 115+) |

```
        sets; then keep the 1000 best sets for the next
        iteration.
    6.  Using the best 1000 sets from step 5, return to step 3
        provided less than 4500 iterations have been done.
    7.  Save resulting best set of values from run j of time
        period i.
  }
  The best set of values from the five sets resulting from
  the preceding loop is used as the set of values to be used
  in the economy, thus permitting the available resources and
  population in each state for the next period to be
  determined (needed for the next pass through the time-
  period loop)
}
```

Following the example of Hamada et al., the algorithm was constructed to have decreasing probabilities of elements being switched as well as decreasing probabilities of elements being mutated as iterations were performed within the $(i, j)$ loops, with the probabilities being restored to higher values at certain numbers of iterations (2001, pp. 177, 180–181). This simulates "punctuated equilibri[a]," which allow argument values to settle towards at least locally-optimizing values; the periodic abrupt increases in the mutation and cross-over probabilities also allow the search space to be explored more (Hamada et al. 2001, p. 177).

The code implementing the genetic algorithm is included in Appendix B.2. The genetic algorithm was coded and run in MATLAB 7.0.1 using programming based on the work of others, most notably Hamada et al. (2001) and Haupt and Haupt (2004). While the genetic algorithm uses different parameters to control mating, crossing-over, and mutation, these parameters are not discussed. Finding appropriate values seems to often require a measured trial-and-error approach, but these particular parameters do not carry much actual meaning.

The simplified version of the extended model allows the constraints to be imposed in the genetic algorithm one time period at a time, beginning with period $t = 0$. The constraints can be imposed sequentially because $\beta = 0$ means that only utility in

the current period matters when determining the amounts to devote to health spending and consumption spending in that period. Consideration of one period at a time in a sequential fashion permits a given period's amount of available resources and state-specific populations to be known prior to applying the algorithm to the given period (based on the argument values and the number of individuals in each health state in the previous period). To force the income constraint to be met, the genetic algorithm calculates the total available income as well as the total income implied by a set of generated arguments. The algorithm scales each individual argument by multiplying each of them by the ratio of available income to argument-implied income, forcing total consumption and health expenditures to equal total income in the period.

Also, it can be shown that in any given period, all individuals would have an identical level of consumption if the objective function were being optimized. Therefore, only one consumption argument for each time period is included so as to improve the efficiency of the algorithm. Thus, each time period has 49 arguments over which to optimize: a common per-capita consumption value, a health spending value for individuals in the "Healthy" risk group for each of the 24 age groups (i.e., 0–4, 5–9, ..., 115+) and a health spending value for individuals in the "Sick" risk group for each of the age groups. The simplification still leaves a substantial number of arguments to be optimally selected.

While convergence criteria were not predetermined, a reasonable retrospective criterion is to determine if all five different sets from a given time period are close to each other. Because these sets all start from randomly-generated values, similarity of the end-values is an indicator that the algorithm might be working. Of course, this is not a sufficient condition to guarantee convergence, even if many more than five runs were made for each time period. Nonetheless, confidence in the argument value results is gained when each run yields similar "optimizing" values.

For each time period, the best set from each of the five runs are compared to see which of these sets yields the highest value for the function to be maximized. Then the argument values in this set are taken to be the actual values for that time period. In so doing, the value of the function to be maximized is determined for that time period, as well as the health levels for all age and risk-factor group combinations and the number of people alive in each state in the next period. Close-to-optimizing arguments for the next period are determined. The process of finding close-to-optimizing arguments and using the arguments to determine the quantities of interest for the next period's optimization is continued until argument values are selected for time period $t = 19$.

# 4. PEDAGOGICAL ANALYSIS OF RESULTS

While the particular results for the examples described in the previous chapter are not important, a brief summary of these results is now included to demonstrate methods that could be used if the necessary data were to be obtained. Recall also that the code producing all output in this section is included in Appendix B.1 and Appendix B.2.

## 4.1 Classification Analysis Example: *The Federalist Papers*

From the classification analysis, the results generated by SAS indicate the performance of the linear classification rule by showing the cross-validation classification error rate. Note that this output is in no way derived from any of the observation records with "Unknown" for the author. Rather, at this stage the parameters used in the linear classification rule have been estimated using all papers with the author known. These parameter estimates are used in the next step—attributing the other papers to either Madison or Hamilton. One helpful indicator as to how this rule might perform is to examine the misclassification (or error) rate using cross-validation. For the 65 papers with the author being identified, Table 4.1 and Table 4.2 summarize how well the linear-classification technique performed on the *The Federalist Papers* data:

Most papers written by Hamilton were assigned to Hamilton, but when Madison was the author, the rule seemed more erratic. The overall error rate was estimated to be 12.3 percent. Also of particular importance is how the papers with an "Unknown" author were classified. Table 4.3 contains the number of "Unknown" papers assigned by the rule to each author, which indicates that seven of them were attributed by the linear classification technique to Madison.

Table 4.1: Classification of *The Federalist Papers* from known author using linear classification technique with prior probabilities

| From author | Number and percent classified into author | | |
|---|---|---|---|
| | Hamilton | Madison | Total |
| Hamilton | 48 | 3 | 51 |
| | 94.12 | 5.88 | 100.00 |
| Madison | 5 | 9 | 14 |
| | 35.71 | 64.29 | 100.00 |
| Total | 53 | 12 | 65 |
| | 81.54 | 18.46 | 100.00 |
| Priors | 0.78462 | 0.21538 | |

Table 4.2: Estimated misclassification rates for known *The Federalist Papers* by author

| | Author | | |
|---|---|---|---|
| | Hamilton | Madison | Total |
| Rate | 0.0588 | 0.3571 | 0.1231 |
| Priors | 0.7846 | 0.2154 | |

Table 4.3: Classification of papers with author "Unknown" using linear classification technique with prior probabilities

| | Number and percent classified into author | | |
|---|---|---|---|
| | Hamilton | Madison | Total |
| From "Unknown" author | 5 | 7 | 12 |
| | 41.67 | 58.33 | 100.00 |
| Prior probabilities | 0.78462 | 0.21538 | |

Table 4.4: Baseline-category logit parameter estimates when baseline category is non-sacrifice out and other categories are combined

| Parameter | Category | DF | Estimate | Standard error | Wald $\chi^2$ | Pr> $\chi^2$ |
|---|---|---|---|---|---|---|
| Intercept | Hit, walk, or sacrifice out | 1 | -0.8422 | 0.3320 | 6.4357 | 0.0112 |
| ERA | Hit, walk, or sacrifice out | 1 | 0.1317 | 0.0787 | 2.8049 | 0.0940 |

## 4.2 Baseline-Category Logit Models Example: Dale Murphy Baseball Data, 1987 Season

From the fitting of a baseline-category logit model in SAS, the parameter estimates are readily obtained. For example, the estimated coefficients when fitting the baseline-category logit model to Dale Murphy's plate appearance data, with getting out as the baseline category and not getting out as the other category, are given in Table 4.4, as obtained from SAS output. This is meant to be analogous in the economic model to the baseline-category logit model with not living to the next period being the baseline category and being alive in the next period (regardless of what health state the person is in for the next period) as the other category.

Similarly, the estimated coefficients from the second baseline-category logit model are included in Table 4.5. Recall that this second model is for all observations where the outcome was not a non-sacrifice out. Here the baseline category has been chosen as either a sacrifice out or a walk. This baseline-category logit model is to estimate the probability of each plate appearance outcome, conditional on it not being an out. This is analogous to estimating the probability of being in each health state in the next period, conditional on the person living to the next health state.

Recall that the goal of this example was to demonstrate that the estimates for

Table 4.5: Baseline-category logit parameter estimates when baseline category is (sacrifice or walk), conditional on not non-sacrifice out. Observations resulting in non-sacrifice outs are excluded in this model.

| Parameter | Category | DF | Estimate | Standard error | Wald $\chi^2$ | Pr> $\chi^2$ |
|---|---|---|---|---|---|---|
| Intercept | Multiple-base hit | 1 | 0.1009 | 0.6258 | 0.0260 | 0.8719 |
| Intercept | Single-base hit | 1 | 0.1043 | 0.5522 | 0.0357 | 0.8501 |
| ERA | Multiple-base hit | 1 | -0.1532 | 0.1474 | 1.0809 | 0.2985 |
| ERA | Single-base hit | 1 | -0.0891 | 0.1281 | 0.4843 | 0.4865 |

the state-to-state transition probability parameters could be easily obtained if the appropriate data were accessible, and, of course, if the model is valid. Once this is done, the probabilities as a function of health spending could be computed. The probability of each plate appearance outcome for Dale Murphy's 1987 baseball plate appearances has been estimated from the aforementioned parameter estimates. A graphical summary of the estimated probabilities is included in Figure 4.1.

One cause for worry is that in this particular example, the model might not have provided a very good fit. Of course, the present goal is not to find a good model for Murphy's baseball performance; any inadequacy in the provided example is not of direct relevance and so does not mean that the state-to-state transition probabilities cannot be modeled well by the proposed relations. Nonetheless, it should be stated that the proposed transition probability model is not necessarily correct. It is important to recognize that if these probabilities are not modeled well by the functions in Section 3.1.2, then alternative state-to-state transition probability relations might need to be implemented.

## 4.3 Genetic Algorithm Example: Implemented with Arbitrary Values

Some pseudo-code for the genetic algorithm has already been included, and the actual code is in Appendix B.2. The resulting argument values from the genetic algorithm are contained in Appendix C for selected time periods. Each time period has five sets of resulting argument values (one for each run) and the population for each health state during that time period. Appendix C suggests the arrival of the resulting argument values close to the optimizing argument values because of the similarity between each of the five sets of resulting argument values for each reported time period. The relationship between the population and the argument values is particularly important, because typically the smaller the percentage of people in a particular health state, the greater the disparity between the corresponding

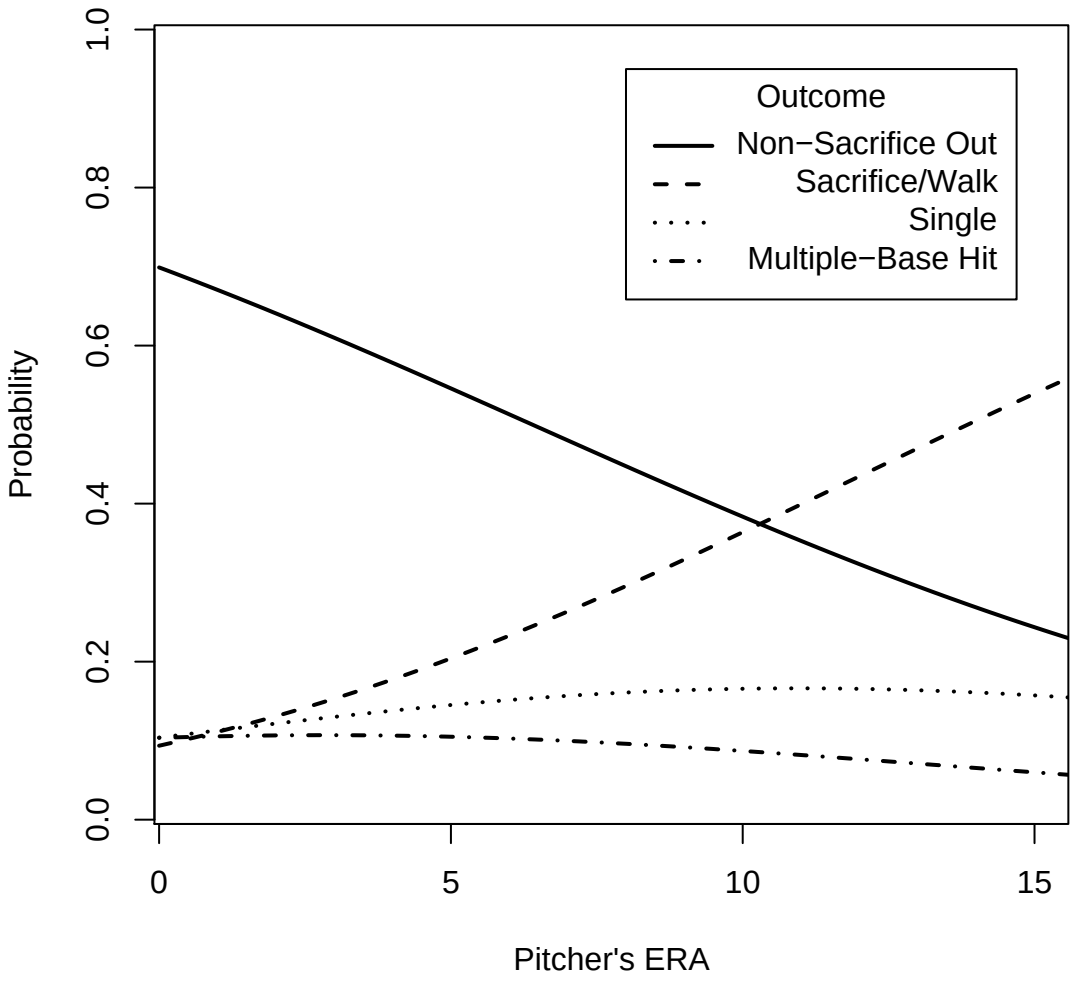


Figure 4.1: Model-estimated probabilities of plate appearance outcomes as a function of ERA for Dale Murphy, 1987 MLB Season

argument values from each of the runs. Even so, the argument values from the five runs are usually relatively close to each other. It is important to note also that in the preliminary time periods, some states had no members; values corresponding to states with no members during that time period cannot have any effect on the overall utility in that period, so the fact that argument values for these states are not similar across runs is not a problem.

One problem that the genetic algorithm might have is that of inaccuracies due to round-off error. Computers do not exactly represent all numbers. In the present case, the argument values from the genetic algorithm saved by MATLAB seem to have lost some precision, at least in the values saved to output files. This is apparent by noting that the income derived from the relation for $y_t$ and the income implied by the argument values written by MATLAB to a file are not exactly the same. However, the order of the relative difference is very small, as the maximum relative difference over the twenty periods is less than 0.00001. The effect of the constraints not being perfectly met is not large enough to detract notably from the implied argument values. Rather, the genetic algorithm's failure to guarantee arrival at the best set of argument values seems to be a more important issue from a practical standpoint, though again, this does not seem to be a major problem in the present case.

From the sets of argument values in Appendix C, it is easy to compute the implied proportion of all income devoted to health spending for each time period. The per-capita consumption for a given time period is already included as an argument, so coupled with the total population in that time period, the total consumption in that period is easily obtained. The total health spending in a given time period can be obtained by summing the total health spending in each health state for the given time period; the total health spending in a given health state for the given time period is the product of the number of people in the given state during the time period and the per-capita health spending of people in that state for that time period.

Then, the total income can be taken as the sum of the total health spending and the total consumption, after which the proportion of the total income used for health spending can be readily obtained. While the total income in time period $t$ could also be obtained by using the relation $\exp(0.02284 \times 5t)$ to determine the per-capita income and multiplying by the total population in period $t$, this was not done because the constraint that total consumption plus total health spending equals total income is not exactly met as a result, probably due to round-off error.

While it is not of direct significance because of the arbitrary nature of the optimization setup, a plot of the near-optimizing health share proportion of income over the time periods is provided in Figure 4.2.

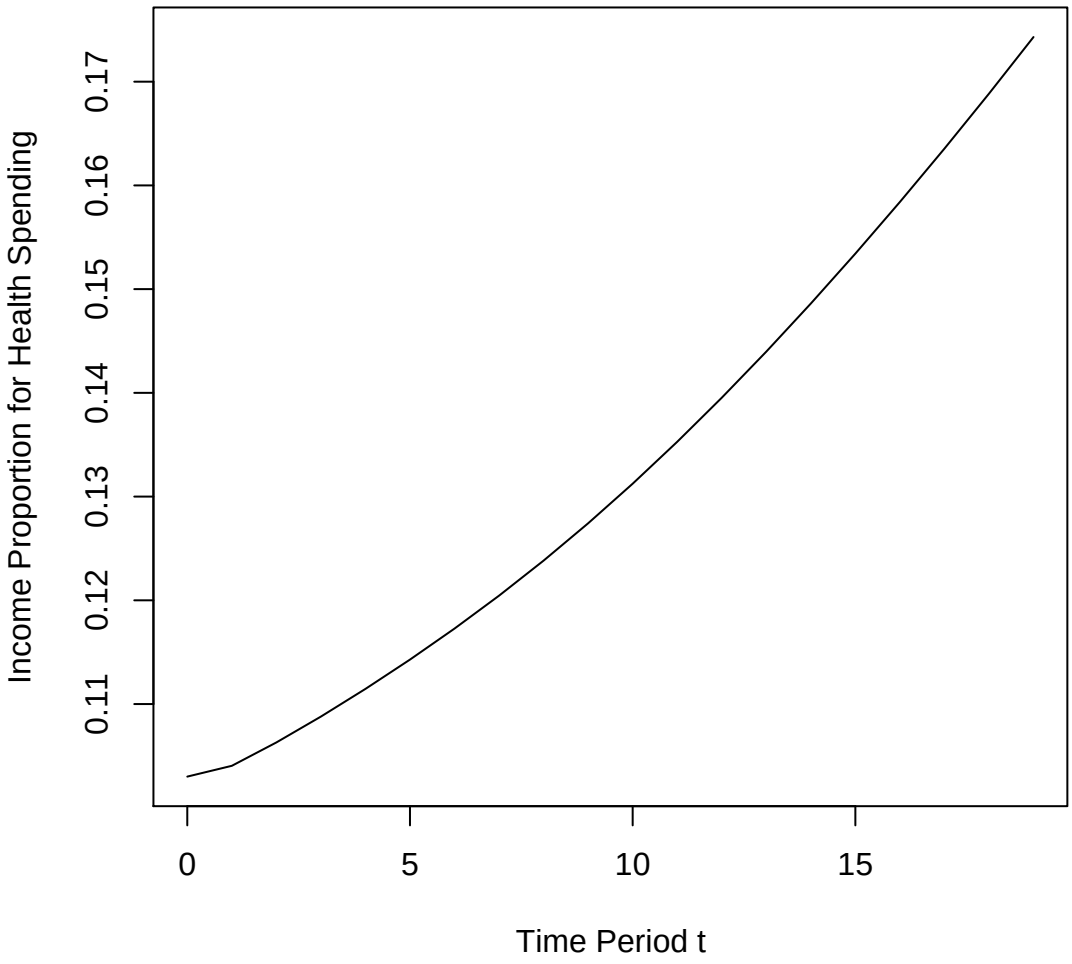


Figure 4.2: Genetic-algorithm-implied income proportion devoted to health spending by period for the (often arbitrary) values used for the optimization setup. Since the values used for this genetic algorithm are often *ad hoc*, these numbers should not be compared with the results by Hall and Jones (2007b). Every time period represents five years, and time period $t = 0$ corresponds to 2005–2009.

# 5. FINAL COMMENTS

This paper has shown how the theoretical model by Hall and Jones (2004b) can be extended to have more representative agents, thus providing for more heterogeneity. It should be noted that the health level equation in the extended theoretical model is a modification which imposes a given form and is thus less general for the theoretical model, but not the empirical model. In addition, a few of the practical tools that could be used in fitting the extended model have been demonstrated briefly; namely, the use of linear classification analysis, the fitting of successive baseline-category logit models, and the use of a genetic algorithm.

In some ways, the "extended" model was not a true extension, as it made a noteworthy change to the relations by which state-to-state transition probabilities are determined. Also, all of the demonstrated tools used in fitting the model utilized approaches different from Hall and Jones (2004b). The viability of the theoretical model's application to the United States is certainly questionable. Complicating the investigation of its viability is the issue of not having the appropriate data to see how well it might fit. However, the main result of this paper is an interesting theoretical extension which can allow for a less stringent model in several aspects. As more and more risk-factor groups are included, the model can allow for an increasingly diverse populace. While this model may or may not be useful in practice, the means by which the theoretical model of Hall and Jones (2004b) was modified might have applications in other macroeconomic models.

At many points in this paper, alternative methods to the procedures in this paper have been highlighted, primarily in the literature review. Alternative ways to deal with the assumption of a representative agent, alternative ways of classifying individuals into health states given data on individuals and data on reference groups,

and alternative ways of fitting a model with a small number of possible categorical outcomes are all detailed in that review.

# A. DATA SOURCES

## A.1 Data on *The Federalist Papers*

The data were downloaded on July 21, 2005 from `http://lib.stat.cmu.edu/datasets/federalistpapers.txt`. The data were submitted by Jeff Collins. I was made aware of the data through StatLib: `http://lib.stat.cmu.edu/datasets/`. Note that Collins also provides links to extra information in his comments on the following web page: `http://lib.stat.cmu.edu/datasets/federalistpapers.txt`.

## A.2 Baseball Data

The data were downloaded on or shortly before April 11, 2005 from `http://www.dartmouth.edu/~chance/teaching_aids/data/baseball.zip`. The data were submitted by Chris Albright. Though the StatLib link to the data appears to no longer be available, I was made aware of the data through the StatLib website: `http://lib.stat.cmu.edu/modules.php?op=modload&name=PostWrap&file=index&page=/DASL`.

## B.1 SAS Code

### Listing B.1: SAS Code to Analyze *The Federalist Papers* Data

```
data fed_Papers;
 infile "F:\FedPapersData.txt" delimiter='09'x MISSOVER DSD lrecl=32767 ;
 input TextNumber        TextName : $12.          Group    FirstPerson      InnerThinking
   ThinkPositive         ThinkNegative    ThinkingAhead    ThinkingBack     WordPicture
   SpaceInterval         Motion   PastEvents      ShiftingEvts     TimeInterval     CueComKnow
   CuePriorText CueReader        CueNotifier      CueMovement      CueReasoning;
run;
proc print data=fed_papers; run;
proc freq; table group; run;

*Split data into two data sets: fed_papers_certain for when
 the author is undisputed, and fed_papers_uncertain for when the
 author is disputed.;
data fed_papers_certain fed_papers_uncertain;
 set fed_papers;
 if (group eq 1) then Author="Hamilton";
 if (group eq 2) then Author="Madison";
 if (group eq 3) then Author="Unknown";
 if (group eq 3) then output fed_papers_uncertain;
 else output fed_papers_certain;
run;

* Change disputed author, or group=3, to missing value.;
data fed_papers_uncertain;
 set fed_papers_uncertain;
 group=.;
 Author="";
run;


* Note that the remainder of the code is based HEAVILY on an example
 given in the on-line SAS documentation for Proc Discrim---specifically
 in Example 25.4 of the documentation for the Discrim Procedure in the
 SAS/STAT section;
*Classify values that were group=3 but are considered as if group is missing;
ods rtf body="F:\federalist_output.rtf";
proc discrim data=fed_papers_certain outstat=rule_info
 method=normal pool=yes crossvalidate;
 class Author;
 var FirstPerson         InnerThinking
   ThinkPositive ThinkNegative ThinkingAhead ThinkingBack WordPicture
   SpaceInterval Motion PastEvents       ShiftingEvts TimeInterval CueComKnow
   CuePriorText CueReader CueNotifier CueMovement CueReasoning;
 priors proportional;
run;


proc discrim data=rule_info testdata=fed_papers_uncertain
 testout=uncertain_classified;
 class Author;
 var FirstPerson         InnerThinking
   ThinkPositive ThinkNegative ThinkingAhead ThinkingBack WordPicture
   SpaceInterval Motion PastEvents       ShiftingEvts TimeInterval CueComKnow
   CuePriorText CueReader CueNotifier CueMovement CueReasoning;
run;
ods rtf close;
proc print data=uncertain_classified; run;

*Seven of the twelve papers with disputed author were classified as being
 from Madison---the other five were classified as being from Hamilton;
```

### Listing B.2: SAS Code to Analyze Murphy Baseball Data

```
data murphy; infile "F:\projectapr27\Project\87murpd1.txt";
  input AtBat I7 O2 HitVal1 Score R123 R23 Success1 Success2 Game
   DN HA T ERA Turf Bats HitVal2 Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9;
run;

data murphy;
  set murphy;
  nonsacout="(non-sacrifice) out";
  nonoutoutcome="multiple-base hit";
  if hitval2 eq 0 then nonoutoutcome="";
  if hitval2 ne 0 then nonsacout="hit, walk, or sac";
  if hitval2 eq 1 then nonoutoutcome="single-base hit";
```

```
  if hitval2 eq 2 then nonoutoutcome="multiple-base hit";
  if hitval2 eq 3 then nonoutoutcome="sacrifice or walk";
  label nonsacout="Out status";
  label nonoutoutcome="Outcome status given not (non-sac) out";
run;

ods html;
ods graphics on;
proc logistic data=murphy;
  model nonsacout(ref="(non-sacrifice) out")=ERA/link=GLOGIT lackfit;
  output out=preds p=predvals;
  graphics estprob;
run;

proc logistic data=murphy;
  model nonoutoutcome(ref="sacrifice or walk")=ERA/link=GLOGIT;
  where nonsacout ne "(non-sacrifice) out";
  output out=preds p=predvals;
  graphics estprob;
run;
ods graphics off;
ods html close;
```

## B.2 MATLAB Code

Listing B.3: MATLAB Code Implementing Genetic Algorithm: Code to Run at Command Line

```
[v11 v22]=main_GA_function(5,20,0.02284,0.01,...
1000000 .* [10 9.95 9.90 9.85 9.80 9.75 ...
       9.70 9.65 9.60 9.55 9.45 9.10 8.50 7.50 6.00 2.00 1.00 ...
       0.20 0.12 0.08 0.01 0.001 0 0 1 1 1 1 1 1.3 1.5 1.6 1.7 ...
       1.8 2.0 2.05 2.00 2.5 2.5 2.0 0.5 0.5 0.5 0.3 0.1 0.002 0.0001 0]);
```

Listing B.4: MATLAB Code Implementing Genetic Algorithm: command_line_code.m

```
%Command Line Code;
function best_vals = command_line_code(t, y_t, n_k_t, z_t)
 J=5;
 best_args=zeros(49,J);

for j=1:J

% Order of values is 48 health spendings in following order:
% healthy/0-4, healthy/5-9, ... healthy/115+, sick/0-4,
% sick/5-9, ..., sick/115+, and then the consumption variable;

%run genetic algorithm, initially just for one time period;

% Let m denote number to generate--at each step there will be this many;
m=3000;

% Create initial pool of argument values;
arg_vals=20 .* rand(49,m);

%Make sure satisfy restraints
implied_income=(n_k_t)' * (arg_vals(1:48,:)) + sum(n_k_t)*arg_vals(49,:);
available_income=sum(n_k_t)*y_t;
arg_vals=arg_vals .* kron((available_income ./ implied_income), ones(49,1));


% Compute health levels;
health_levels=compute_health(arg_vals(1:48,:),z_t,m);

%Compute obj_function;
obj_value=compute_obj_function(n_k_t(:,1),arg_vals(49,:),health_levels,m);

% rank arg_vals sets;
[s_o_v rank_index]=sort(obj_value,'descend');
rank_arg_vals=arg_vals(:,rank_index);

%Now for GA work: have initial pool with ranked values;
gen_num=0;
G=4500;

for i=1:G
    %mate;
    %create function to find pairs to mate;
    obs_to_mate=mate_selection(m);
    %create function to perform cross-overs;
    cross_over_result=cross_over(rank_arg_vals(:,obs_to_mate));
    %create function to perform mutations;
    mutate_result=mutate(rank_arg_vals(:,1:ceil(m/3)),gen_num);
```

```
    %Combine original good, cross-over, and mutated into arg_vals;
    arg_vals(:,1:ceil(m/3))=rank_arg_vals(:,1:ceil(m/3));
    arg_vals(:,(m/3 +1):(2*m/3))=cross_over_result;
    arg_vals(:,(2*m/3 +1):m)=mutate_result;


    % make sure satisfy constraints;
    implied_income=(n_k_t)' * (arg_vals(1:48,:)) + sum(n_k_t)*arg_vals(49,:);
    available_income=sum(n_k_t)*y_t;
    arg_vals=arg_vals .* kron((available_income ./ implied_income), ones(49,1));

    %call functions compute_health, compute_obj_function, and sort by rank;
    health_levels=compute_health(arg_vals(1:48,:),z_t,m);
    obj_value=compute_obj_function(n_k_t,arg_vals(49,:),health_levels,m);


    [s_o_v rank_index]=sort(obj_value,'descend');
    rank_arg_vals=arg_vals(:,rank_index);
    gen_num=gen_num+1;

end

best_args(:,j)=rank_arg_vals(:,1);


end


file=['Best_4_period_' num2str(t)];
save(file, 'best_args', '-ascii');

    best_health_levels=compute_health(best_args(1:48,:),z_t,J);
    best_obj_values= ...
      compute_obj_function(n_k_t,best_args(49,:),best_health_levels,J);


    [s_b_o_v rank_index_best]=sort(best_obj_values,'descend');
    ranked_best_vals=best_args(:,rank_index_best);

    best_vals=ranked_best_vals(:,1);

end
```

Listing B.5: MATLAB Code Implementing Genetic Algorithm: compute_health.m

```
function h_level=compute_health(health_exp, tech_level,num)
  alpha_live=[5.2   4.9   4.6   4.3   4.0   3.7   3.4   3.1 ...
              2.8   2.5   2.2   1.9   1.6   1.3   1.0   0.7 ...
              0.4   0.1  -0.2  -0.5  -0.8  -1.1  -1.4  -1.7 ...
              4.2   3.9   3.6   3.3   3.0   2.7   2.4   2.1 ...
              1.8   1.5   1.2   0.9   0.6   0.3   0.0  -0.3 ...
             -0.6  -0.9  -1.2  -1.5  -1.8  -2.1  -2.4  -2.7];

  alpha_h_given_l=alpha_live;

  beta_live=[0.48  0.46  0.44  0.42  0.40  0.38  0.36  0.34 ...
             0.32  0.30  0.28  0.26  0.24  0.22  0.20  0.18 ...
             0.16  0.14  0.12  0.10  0.08  0.06  0.04  0.02 ...
             0.48  0.46  0.44  0.42  0.40  0.38  0.36  0.34 ...
             0.32  0.30  0.28  0.26  0.24  0.22  0.20  0.18 ...
             0.16  0.14  0.12  0.10  0.08  0.06  0.04  0.02];

  beta_h_given_l=beta_live;
  ehe=health_exp .* tech_level;
  lehe=log(ehe);
  %compute odds of living;%It equals exp(alpha+beta*ln(eff.health.expend));
  odds_live=exp(kron(alpha_live',ones(1,num)) + ...
   (kron(beta_live',ones(1,num)) .* lehe));

  prob_live=odds_live ./ (ones(48,num)+odds_live);

  odds_h_live=exp(kron(alpha_h_given_l',ones(1,num)) + ...
   (kron(beta_h_given_l',ones(1,num)) .* lehe));

  prob_h_live=odds_h_live ./(ones(48,num)+odds_h_live);

  prob_healthy= prob_live .* prob_h_live;
  prob_sick=prob_live .* (1-prob_h_live);

  h_level=(1 ./ (1-prob_healthy- (0.7 .* prob_sick)));

end
```

Listing B.6: MATLAB Code Implementing Genetic Algorithm: compute_obj_function.m

```
function total_utility=compute_obj_function(N_vec,cons_lev,health_lev,num)
```

```
  utility_per_person= ((((kron((cons_lev), ones(48,1)  ).^ (-0.6)) ...
       ./ (-0.6) )) + (((health_lev) .^ (-0.7)) .* (2/(-0.7))) ...
                              + (15 .* ones(48,num)));

  total_utility=N_vec' * utility_per_person;
end
```

Listing B.7: MATLAB Code Implementing Genetic Algorithm: mate_selection.m

```
function col_nums=mate_selection(num)
  num_resulting=ceil(num/3);
  the_sum=sum(1:num_resulting);
  init_draws=ceil(rand(num_resulting,1) .* the_sum);


  col_nums=ceil((-2*num_resulting-1 ...
      +sqrt((2*num_resulting+1)*(2*num_resulting+1)-8*init_draws)) ./ (-2));
end
```

Listing B.8: MATLAB Code Implementing Genetic Algorithm: cross_over.m

```
function cross_over_results=cross_over(initial_vals)
    pairs2switch= ...
        (floor(rand(size(initial_vals,1), ...
        (size(initial_vals,2) ./ 2)) .* (10 ./ 6.5)));
    make_switch_1=kron(pairs2switch,[1 0]);
    make_switch_2=kron(pairs2switch,[0 1]);



    temp_vals=initial_vals;
  m_s_log_1=make_switch_1 > 0;
  m_s_log_2=make_switch_2 > 0;
  temp_vals(m_s_log_1)=initial_vals(m_s_log_2);
  temp_vals(m_s_log_2)=initial_vals(m_s_log_1);
  cross_over_results=temp_vals;
end
```

Listing B.9: MATLAB Code Implementing Genetic Algorithm: mutate.m

```
function output_vals=mutate(input_vals, the_generation)
%mutate function to simulate mutation of argument values
%  I designed it to have decreasing probability of mutation, as well
%  as decreasing variance of mutation, but with periodic resets
%  (like the punctuated equilibrium in Hamada et al 2001)
%
%  The period is 300 generations
%
%  The inputs are a matrix of original values for
%  the (m/3, or population size/3) best argument sets and the generation
%  number, which is needed to specify how the mutation should behave
%  in that period.
%
%  The output is the matrix of mutated values.  This will be appended to
%  the original values;


  vals2mutate= ...
   floor(rand(size(input_vals,1),size(input_vals,2)) ...
   .* (10 ./ (log(1096.633 + mod(the_generation,300) .* 13668.15 ./ 299))));
  vals2mutate_log = vals2mutate > 0;

  % The mutation rate varies from about 30 percent to about 4 percent;


  % the mutation proportion becomes more and more concentrated at 1
  % until the reset of the the alpha and beta values;
  mutate_prop=ones(size(input_vals,1),size(input_vals,2));
  mutate_prop_4_dev= 2 .* betarnd((10 + 2 .* mod(the_generation,300)), ...
      (10 + 2 .* mod(the_generation,300)), ...
      size(input_vals,1),size(input_vals,2));

  mutate_prop(vals2mutate_log)=mutate_prop_4_dev(vals2mutate_log);

  output_vals=input_vals .* mutate_prop;
end
```

Listing B.10: MATLAB Code Implementing Genetic Algorithm: get_new_pop_count.m

```
function the_pop_in_t = get_new_pop_count(prev_pop,log_eff_h_inputs)
    init_pop=prev_pop;

    % here I compute the transition probabilities;
```

```
alpha_live=[5.2   4.9   4.6   4.3   4.0   3.7   3.4   3.1 ...
            2.8   2.5   2.2   1.9   1.6   1.3   1.0   0.7 ...
            0.4   0.1  -0.2  -0.5  -0.8  -1.1  -1.4  -1.7 ...
            4.2   3.9   3.6   3.3   3.0   2.7   2.4   2.1 ...
            1.8   1.5   1.2   0.9   0.6   0.3   0.0  -0.3 ...
           -0.6  -0.9  -1.2  -1.5  -1.8  -2.1  -2.4  -2.7];

alpha_h_given_l=alpha_live;

beta_live=[0.48  0.46  0.44  0.42  0.40  0.38  0.36  0.34 ...
           0.32  0.30  0.28  0.26  0.24  0.22  0.20  0.18 ...
           0.16  0.14  0.12  0.10  0.08  0.06  0.04  0.02 ...
           0.48  0.46  0.44  0.42  0.40  0.38  0.36  0.34 ...
           0.32  0.30  0.28  0.26  0.24  0.22  0.20  0.18 ...
           0.16  0.14  0.12  0.10  0.08  0.06  0.04  0.02];

beta_h_given_l=beta_live;

%compute odds of living;%It equals exp(alpha+beta*ln(eff.health.expend));
odds_live=exp(alpha_live' + (beta_live' .* log_eff_h_inputs));

prob_live=odds_live ./ (ones(48,1)+odds_live);

odds_h_live=exp(alpha_h_given_l' + (beta_h_given_l' .* log_eff_h_inputs));

prob_h_live=odds_h_live ./ (ones(48,1)+odds_h_live);

prob_healthy= prob_live .* prob_h_live;
prob_sick=prob_live .* (1-prob_h_live);


% These next two variables do not take into account age category
% generally being incremented;

num_healthy_next_period=init_pop(1:24,1) .* prob_healthy(1:24,1) + ...
    init_pop(25:48,1) .* prob_healthy(25:48,1);

num_sick_next_period=init_pop(1:24,1) .* prob_sick(1:24,1) + ...
    init_pop(25:48,1) .* prob_sick(25:48,1);


% Now I adjust for age usually going up, with 115+ staying in 115+
% age group if survive, and also constant amount of people in each
% initial age group, though numbers differ across risk groups.
n_h_n_p=[10000000 (num_healthy_next_period(1:22,1)') ...
    (num_healthy_next_period(23,1)+num_healthy_next_period(24,1))];


n_s_n_p=[1000000 (num_sick_next_period(1:22,1)') ...
    (num_sick_next_period(23,1)+num_sick_next_period(24,1))];

%Now put it all together for output;
the_pop_in_t_row=[n_h_n_p n_s_n_p];
the_pop_in_t=the_pop_in_t_row';

end
```

Listing B.11: MATLAB Code to Rank Values and Save Best Run for Each Period: usefulMatlab

```
% Import all the data

% get obj_function_value for each run


obj_function_values=zeros(20,5)

k=0; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_1(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_1(49,:),h_lev,5)

k=1; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_2(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_2(49,:),h_lev,5)

k=2; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_3(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_3(49,:),h_lev,5)

k=3; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_4(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_4(49,:),h_lev,5)

k=4; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_5(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_5(49,:),h_lev,5)

k=5; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_6(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_6(49,:),h_lev,5)

k=6; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_7(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_7(49,:),h_lev,5)

k=7; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_8(1:48,:),z_t,5);
```

```
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_8(49,:),h_lev,5)

k=8; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_9(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_9(49,:),h_lev,5)

k=9; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_10(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_10(49,:),h_lev,5)

k=10; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_11(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_11(49,:),h_lev,5)

k=11; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_12(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_12(49,:),h_lev,5)

k=12; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_13(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_13(49,:),h_lev,5)

k=13; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_14(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_14(49,:),h_lev,5)

k=14; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_15(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_15(49,:),h_lev,5)

k=15; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_16(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_16(49,:),h_lev,5)

k=16; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_17(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_17(49,:),h_lev,5)

k=17; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_18(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_18(49,:),h_lev,5)

k=18; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_19(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_19(49,:),h_lev,5)

k=19; z_t=exp(0.05*k); h_lev=compute_health(Best_4_period_20(1:48,:),z_t,5);
obj_function_values((k+1),:)=compute_obj_function(for_v22(:,(k+1)),Best_4_period_20(49,:),h_lev,5)

% Now have objective function values
```

## B.3 R Code

### Listing B.12: R Code to Make Baseball Plot

```
# These estimated values were obtained from SAS
fit1coef<- c(-0.8422, 0.1317)

fit2coef<- matrix(c(0.1009, 0.1043,-0.1532, -0.0891),nrow=2,byrow=FALSE)

out.prob <- function(x,coef1,coef2) {1/(1+exp(coef1+coef2*x))}

which.prob<- function(x,int.prob,slope.prob,other.int1,other.slope1,
 other.int2,other.slope2,non.out.int,non.out.slope) {
   (exp(non.out.int + non.out.slope*x)/(exp(non.out.int + non.out.slope *x)+1)) *
   (exp(int.prob + slope.prob*x)/(exp(int.prob + slope.prob*x) +
    exp(other.int1 + other.slope1*x) + exp(other.int2 + other.slope2*x)))
}
tt<-seq(0,20,by=0.05)

par(mfcol=c(1,1))
pdf(file="F:/projectapr27/Project/Murphy.pdf", height=6, width=6)

# Plot of probability of getting out as a function of ERA
plot(tt,out.prob(tt,fit1coef[1],fit1coef[2]),type="l",ylim=c(0.032,0.968),
 xlab="Pitcher's ERA",
 ylab="Probability",lwd=2,xlim=c(0.5,15)
#main="Model Estimated Probabilities of Plate
#Appearance Outcomes as a Function of ERA\nfor Dale Murphy, 1987"
)

# Add to plot probability of sacrifice or walk as function of ERA
lines(tt,which.prob(tt,0,0,fit2coef[1,1],fit2coef[1,2],fit2coef[2,1],fit2coef[2,2],
 fit1coef[1],fit1coef[2]),lty=2,col="black",lwd=2)

# Add to plot probability of single as a function of ERA
lines(tt,which.prob(tt,fit2coef[2,1],fit2coef[2,2],0,0,fit2coef[1,1],fit2coef[1,2],
 fit1coef[1],fit1coef[2]),lty=3,col="black",lwd=2)

# Add to plot probability of multiple-base hit as a function of ERA
lines(tt,which.prob(tt,fit2coef[1,1],fit2coef[1,2],0,0,fit2coef[2,1],fit2coef[2,2],
 fit1coef[1],fit1coef[2]),lty=4,col="black",lwd=2)

# Add a legend: modified code from ?legend
temp <- legend(8,0.95, legend = c(" "," "," "," "), lty=1:4, title="Outcome",
 col=c("black","black","black","black"), text.width=strwidth("Non-Sacrifice Out"),
 lwd=2)
text(temp$rect$left + temp$rect$w, temp$text$y,
```

```
 c("Non-Sacrifice Out", "Sacrifice/Walk", "Single", "Multiple-Base Hit"), pos=2)
dev.off()
```

## Listing B.13: R Code to Get Populations Numbers

```
# This is code to get summary population numbers for selected time periods

thePopulations<-read.table("E:/for_v22.txt")

round(thePopulations[,c(1,2,6,7,11,12,16,17)])
thePopulations[24,c(1,2,6,7,11,12,16,17)]
apply(thePopulations[,c(1,2,6,7,11,12,16,17)],2,sum)
```

## Listing B.14: R Code to Make Health Proportion Plot

```
# Code to make health spending plot
setwd("E:/tempThesis")
for.v11<-as.matrix(read.table("for_v11"))
for.v22<-as.matrix(read.table("for_v22"))

for.v11
for.v22

pop.sum<-apply(for.v22,2,sum)
total.cons<-pop.sum * for.v11[49,]
health.sp.per.state<-for.v22
total.health<-diag(t(for.v22) %*% for.v11[-49,])

total.income1<-total.cons+total.health
total.income2<-exp(0:19 * (5*0.02284)) *pop.sum
total.income1
total.income2

prop.health1<-total.health/total.income1
prop.health2<-total.health/total.income2

prop.health3<- 1 - (total.cons/total.income1)
prop.health4<- 1- (total.cons/total.income2)
max.percent.diff<-function(x,y) {100*max(abs(2*(x- y)/(x+y)))}
max.percent.diff(total.income1, total.income2)
max.percent.diff(prop.health1,prop.health2)

pdf(file="F:/Project/healthPlot.pdf",height=6,width=6)
par(mfrow=c(1,1))
plot(0:19,prop.health1,type="l",
 main="",
 xlab="Time Period t",
 ylab="Income Proportion for Health Spending")
dev.off()
```

# C. GENETIC ALGORITHM ARGUMENT VALUES FOR SELECTED TIME PERIODS

Table C.1: Information from five genetic algorithm runs for time period 0 (2005–2009)

| For state | Associated argument (of form $h_{g,a,t}$ or $c_t$) | Argument value (per state member) in 2005 per capita income units | | | | | Population for present period | Implied population for next period from best run |
|---|---|---|---|---|---|---|---|---|
| | | Run 1 | Run 2 | Run 3 | Run 4 | Run 5[a] | | |
| Healthy, ages 0–4 | $h_{1,0-4,0}$ | 0.062022 | 0.062022 | 0.062022 | 0.062022 | 0.062022 | 10,000,000 | 10,000,000 |
| Healthy, 5–9 | $h_{1,5-9,0}$ | 0.067858 | 0.067858 | 0.067858 | 0.067858 | 0.067858 | 9,950,000 | 10,509,352 |
| Healthy, 10–14 | $h_{1,10-14,0}$ | 0.074145 | 0.074145 | 0.074145 | 0.074145 | 0.074145 | 9,900,000 | 10,355,187 |
| Healthy, 15–19 | $h_{1,15-19,0}$ | 0.080853 | 0.080853 | 0.080853 | 0.080853 | 0.080853 | 9,850,000 | 10,177,275 |
| Healthy, 20–24 | $h_{1,20-24,0}$ | 0.087916 | 0.087918 | 0.087915 | 0.087915 | 0.087915 | 9,800,000 | 9,970,164 |
| Healthy, 25–29 | $h_{1,25-29,0}$ | 0.095216 | 0.095217 | 0.095217 | 0.095217 | 0.095217 | 9,750,000 | 9,727,321 |
| Healthy, 30–34 | $h_{1,30-34,0}$ | 0.102572 | 0.102574 | 0.102572 | 0.102572 | 0.102572 | 9,700,000 | 9,672,278 |
| Healthy, 35–39 | $h_{1,35-39,0}$ | 0.109702 | 0.109702 | 0.109702 | 0.109702 | 0.109702 | 9,650,000 | 9,463,376 |
| Healthy, 40–44 | $h_{1,40-44,0}$ | 0.116216 | 0.116215 | 0.116215 | 0.116215 | 0.116216 | 9,600,000 | 9,100,850 |
| Healthy, 45–49 | $h_{1,45-49,0}$ | 0.121590 | 0.121590 | 0.121589 | 0.121590 | 0.121590 | 9,550,000 | 8,649,835 |
| Healthy, 50–54 | $h_{1,50-54,0}$ | 0.125175 | 0.125175 | 0.125175 | 0.125175 | 0.125176 | 9,450,000 | 8,100,626 |
| Healthy, 55–59 | $h_{1,55-59,0}$ | 0.126231 | 0.126232 | 0.126232 | 0.126232 | 0.126232 | 9,100,000 | 7,457,642 |
| Healthy, 60–64 | $h_{1,60-64,0}$ | 0.124035 | 0.124035 | 0.124035 | 0.124035 | 0.124035 | 8,500,000 | 6,493,960 |
| Healthy, 65–69 | $h_{1,65-69,0}$ | 0.118048 | 0.118048 | 0.118048 | 0.118048 | 0.118048 | 7,500,000 | 5,331,025 |
| Healthy, 70–74 | $h_{1,70-74,0}$ | 0.108151 | 0.108151 | 0.108151 | 0.108151 | 0.108151 | 6,000,000 | 4,142,320 |
| Healthy, 75–79 | $h_{1,75-79,0}$ | 0.094837 | 0.094837 | 0.094837 | 0.094837 | 0.094836 | 2,000,000 | 2,777,754 |
| Healthy, 80–84 | $h_{1,80-84,0}$ | 0.079244 | 0.079244 | 0.079243 | 0.079244 | 0.079243 | 1,000,000 | 838,954 |
| Healthy, 85–89 | $h_{1,85-89,0}$ | 0.062924 | 0.062924 | 0.062924 | 0.062924 | 0.062923 | 200,000 | 280,106 |
| Healthy, 90–94 | $h_{1,90-94,0}$ | 0.047422 | 0.047421 | 0.047422 | 0.047422 | 0.047422 | 120,000 | 56,981 |
| Healthy, 95–99 | $h_{1,95-99,0}$ | 0.033867 | 0.033867 | 0.033867 | 0.033868 | 0.033867 | 80,000 | 28,602 |
| Healthy, 100–104 | $h_{1,100-104,0}$ | 0.022797 | 0.022798 | 0.022798 | 0.022797 | 0.022798 | 10,000 | 12,193 |
| Healthy, 105–109 | $h_{1,105-109,0}$ | 0.014236 | 0.014236 | 0.014234 | 0.014234 | 0.014235 | 1,000 | 1,669 |
| Healthy, 110–114 | $h_{1,110-114,0}$ | 0.290470 | 0.266179 | 0.627921 | 0.725581 | 0.178767 | 0 | 56 |
| Healthy, 115+ | $h_{1,115+,0}$ | 0.046522 | 0.394526 | 0.196810 | 0.388436 | 0.101274 | 0 | 0[b] |
| Sick, 0–4 | $h_{2,0-4,0}$ | 0.100985 | 0.100986 | 0.100986 | 0.100986 | 0.100986 | 1,000,000 | 1,000,000 |
| Sick, 5–9 | $h_{2,5-9,0}$ | 0.110077 | 0.110078 | 0.110078 | 0.110078 | 0.110078 | 1,000,000 | 242,288 |
| Sick, 10–14 | $h_{2,10-14,0}$ | 0.119534 | 0.119534 | 0.119534 | 0.119532 | 0.119534 | 1,000,000 | 292,891 |
| Sick, 15–19 | $h_{2,15-19,0}$ | 0.129129 | 0.129126 | 0.129128 | 0.129129 | 0.129128 | 1,000,000 | 354,608 |
| Sick, 20–24 | $h_{2,20-24,0}$ | 0.138513 | 0.138513 | 0.138513 | 0.138513 | 0.138513 | 1,000,000 | 429,759 |
| Sick, 25–29 | $h_{2,25-29,0}$ | 0.147188 | 0.147187 | 0.147187 | 0.147188 | 0.147187 | 1,300,000 | 520,988 |
| Sick, 30–34 | $h_{2,30-34,0}$ | 0.154474 | 0.154474 | 0.154474 | 0.154474 | 0.154475 | 1,500,000 | 663,329 |
| Sick, 35–39 | $h_{2,35-39,0}$ | 0.159514 | 0.159514 | 0.159514 | 0.159513 | 0.159514 | 1,600,000 | 827,182 |
| Sick, 40–44 | $h_{2,40-44,0}$ | 0.161317 | 0.161317 | 0.161317 | 0.161317 | 0.161317 | 1,700,000 | 1,010,214 |
| Sick, 45–49 | $h_{2,45-49,0}$ | 0.158894 | 0.158896 | 0.158895 | 0.158897 | 0.158895 | 1,800,000 | 1,224,666 |
| Sick, 50–54 | $h_{2,50-54,0}$ | 0.151511 | 0.151510 | 0.151510 | 0.151510 | 0.151510 | 2,000,000 | 1,469,031 |
| Sick, 55–59 | $h_{2,55-59,0}$ | 0.138991 | 0.138989 | 0.138991 | 0.138990 | 0.138991 | 2,050,000 | 1,752,486 |
| Sick, 60–64 | $h_{2,60-64,0}$ | 0.122026 | 0.122026 | 0.122027 | 0.122027 | 0.122026 | 2,000,000 | 1,971,118 |
| Sick, 65–69 | $h_{2,65-69,0}$ | 0.102207 | 0.102208 | 0.102207 | 0.102209 | 0.102207 | 2,500,000 | 2,092,252 |
| Sick, 70–74 | $h_{2,70-74,0}$ | 0.081663 | 0.081663 | 0.081663 | 0.081665 | 0.081663 | 2,500,000 | 2,204,559 |
| Sick, 75–79 | $h_{2,75-79,0}$ | 0.062430 | 0.062429 | 0.062430 | 0.062430 | 0.062430 | 2,000,000 | 1,977,474 |
| Sick, 80–84 | $h_{2,80-84,0}$ | 0.045895 | 0.045896 | 0.045895 | 0.045895 | 0.045896 | 500,000 | 918,537 |
| Sick, 85–89 | $h_{2,85-89,0}$ | 0.032613 | 0.032612 | 0.032612 | 0.032612 | 0.032612 | 500,000 | 344,010 |
| Sick, 90–94 | $h_{2,90-94,0}$ | 0.022468 | 0.022469 | 0.022469 | 0.022469 | 0.022468 | 500,000 | 129,322 |
| Sick, 95–99 | $h_{2,95-99,0}$ | 0.014997 | 0.014997 | 0.014997 | 0.014997 | 0.014997 | 300,000 | 95,047 |
| Sick, 100–104 | $h_{2,100-104,0}$ | 0.009627 | 0.009627 | 0.009627 | 0.009627 | 0.009627 | 100,000 | 50,313 |
| Sick, 105–109 | $h_{2,105-109,0}$ | 0.005829 | 0.005828 | 0.005829 | 0.005829 | 0.005829 | 2,000 | 11,058 |
| Sick, 110–114 | $h_{2,110-114,0}$ | 0.003167 | 0.003167 | 0.003167 | 0.003167 | 0.003167 | 100 | 314 |
| Sick, 115+ | $h_{2,115+,0}$ | 0.013924 | 0.231224 | 0.260045 | 0.364651 | 0.500093 | 0 | 6 |
| All | $c_{t=0}$ | 0.896995 | 0.896995 | 0.896995 | 0.896995 | 0.896995 | 169,563,100 | 152,728,981 |

NOTE: For display purposes, the population numbers were rounded to the nearest integer using the R function *round*.
[a]Best of the 5 runs.
[b]This number is about 0.45.

Table C.2: Information from five genetic algorithm runs for time period 5 (2030–2034)

| For state | Associated argument (of form $h_{g,a,t}$ or $c_t$) | Argument value (per state member) in 2005 per capita income units | | | | | Population for present period | Implied population for next period from best run |
|---|---|---|---|---|---|---|---|---|
| | | Run 1[a] | Run 2 | Run 3 | Run 4 | Run 5 | | |
| Healthy, ages 0–4 | $h_{1,0-4,5}$ | 0.114921 | 0.114921 | 0.114921 | 0.114921 | 0.114921 | 10,000,000 | 10,000,000 |
| Healthy, 5–9 | $h_{1,5-9,5}$ | 0.127164 | 0.127164 | 0.127164 | 0.127164 | 0.127164 | 10,645,073 | 10,672,602 |
| Healthy, 10–14 | $h_{1,10-14,5}$ | 0.140619 | 0.140620 | 0.140620 | 0.140619 | 0.140619 | 10,407,755 | 10,452,882 |
| Healthy, 15–19 | $h_{1,15-19,5}$ | 0.155311 | 0.155311 | 0.155310 | 0.155311 | 0.155311 | 10,089,877 | 10,157,160 |
| Healthy, 20–24 | $h_{1,20-24,5}$ | 0.171207 | 0.171207 | 0.171207 | 0.171207 | 0.171207 | 9,690,583 | 9,784,053 |
| Healthy, 25–29 | $h_{1,25-29,5}$ | 0.188190 | 0.188191 | 0.188191 | 0.188190 | 0.188190 | 9,195,823 | 9,319,468 |
| Healthy, 30–34 | $h_{1,30-34,5}$ | 0.206013 | 0.206014 | 0.206013 | 0.206015 | 0.206013 | 8,750,355 | 8,748,731 |
| Healthy, 35–39 | $h_{1,35-39,5}$ | 0.224242 | 0.224242 | 0.224242 | 0.224242 | 0.224245 | 8,219,074 | 8,207,642 |
| Healthy, 40–44 | $h_{1,40-44,5}$ | 0.242187 | 0.242188 | 0.242188 | 0.242187 | 0.242187 | 7,588,227 | 7,565,753 |
| Healthy, 45–49 | $h_{1,45-49,5}$ | 0.258830 | 0.258831 | 0.258830 | 0.258830 | 0.258829 | 6,846,470 | 6,811,958 |
| Healthy, 50–54 | $h_{1,50-54,5}$ | 0.272758 | 0.272758 | 0.272755 | 0.272758 | 0.272765 | 6,145,106 | 5,942,266 |
| Healthy, 55–59 | $h_{1,55-59,5}$ | 0.282162 | 0.282159 | 0.282163 | 0.282162 | 0.282163 | 5,240,740 | 5,096,065 |
| Healthy, 60–64 | $h_{1,60-64,5}$ | 0.284955 | 0.284955 | 0.284955 | 0.284955 | 0.284954 | 4,189,205 | 4,086,491 |
| Healthy, 65–69 | $h_{1,65-69,5}$ | 0.279091 | 0.279091 | 0.279090 | 0.279091 | 0.279090 | 3,107,338 | 3,006,560 |
| Healthy, 70–74 | $h_{1,70-74,5}$ | 0.263148 | 0.263151 | 0.263148 | 0.263148 | 0.263147 | 2,091,252 | 1,996,994 |
| Healthy, 75–79 | $h_{1,75-79,5}$ | 0.237060 | 0.237060 | 0.237060 | 0.237061 | 0.237059 | 1,248,035 | 1,163,669 |
| Healthy, 80–84 | $h_{1,80-84,5}$ | 0.202671 | 0.202671 | 0.202671 | 0.202671 | 0.202672 | 621,047 | 578,522 |
| Healthy, 85–89 | $h_{1,85-89,5}$ | 0.163632 | 0.163630 | 0.163632 | 0.163629 | 0.163631 | 252,806 | 230,257 |
| Healthy, 90–94 | $h_{1,90-94,5}$ | 0.124417 | 0.124417 | 0.124418 | 0.124417 | 0.124418 | 84,769 | 72,142 |
| Healthy, 95–99 | $h_{1,95-99,5}$ | 0.088925 | 0.088928 | 0.088927 | 0.088926 | 0.088926 | 21,284 | 18,053 |
| Healthy, 100–104 | $h_{1,100-104,5}$ | 0.059482 | 0.059489 | 0.059486 | 0.059485 | 0.059479 | 2,231 | 3,321 |
| Healthy, 105–109 | $h_{1,105-109,5}$ | 0.036692 | 0.036691 | 0.036697 | 0.036706 | 0.036704 | 219 | 254 |
| Healthy, 110–114 | $h_{1,110-114,5}$ | 0.019980 | 0.019988 | 0.019981 | 0.019995 | 0.019979 | 15 | 18 |
| Healthy, 115+ | $h_{1,115+,5}$ | 0.008172 | 0.008169 | 0.008169 | 0.008169 | 0.008188 | 3[b] | 1[c] |
| Sick, 0–4 | $h_{2,0-4,5}$ | 0.189451 | 0.189450 | 0.189451 | 0.189451 | 0.189446 | 1,000,000 | 1,000,000 |
| Sick, 5–9 | $h_{2,5-9,5}$ | 0.209432 | 0.209431 | 0.209432 | 0.209430 | 0.209432 | 175,885 | 162,357 |
| Sick, 10–14 | $h_{2,10-14,5}$ | 0.230951 | 0.230950 | 0.230950 | 0.230950 | 0.230951 | 197,093 | 182,414 |
| Sick, 15–19 | $h_{2,15-19,5}$ | 0.253762 | 0.253763 | 0.253761 | 0.253766 | 0.253764 | 238,093 | 221,375 |
| Sick, 20–24 | $h_{2,20-24,5}$ | 0.277400 | 0.277401 | 0.277398 | 0.277398 | 0.277406 | 286,954 | 268,174 |
| Sick, 25–29 | $h_{2,25-29,5}$ | 0.301086 | 0.301086 | 0.301086 | 0.301087 | 0.301091 | 343,824 | 323,220 |
| Sick, 30–34 | $h_{2,30-34,5}$ | 0.323618 | 0.323622 | 0.323617 | 0.323617 | 0.323617 | 415,892 | 386,516 |
| Sick, 35–39 | $h_{2,35-39,5}$ | 0.343290 | 0.343289 | 0.343288 | 0.343292 | 0.343290 | 500,268 | 465,325 |
| Sick, 40–44 | $h_{2,40-44,5}$ | 0.357831 | 0.357832 | 0.357831 | 0.357830 | 0.357839 | 596,325 | 554,920 |
| Sick, 45–49 | $h_{2,45-49,5}$ | 0.364514 | 0.364514 | 0.364512 | 0.364517 | 0.364512 | 700,878 | 652,161 |
| Sick, 50–54 | $h_{2,50-54,5}$ | 0.360510 | 0.360511 | 0.360510 | 0.360507 | 0.360510 | 827,396 | 749,773 |
| Sick, 55–59 | $h_{2,55-59,5}$ | 0.343621 | 0.343620 | 0.343620 | 0.343620 | 0.343623 | 937,528 | 856,152 |
| Sick, 60–64 | $h_{2,60-64,5}$ | 0.313288 | 0.313288 | 0.313291 | 0.313290 | 0.313290 | 1,006,022 | 923,768 |
| Sick, 65–69 | $h_{2,65-69,5}$ | 0.271510 | 0.271511 | 0.271514 | 0.271510 | 0.271510 | 1,011,820 | 923,954 |
| Sick, 70–74 | $h_{2,70-74,5}$ | 0.222887 | 0.222887 | 0.222887 | 0.222887 | 0.222882 | 931,956 | 842,367 |
| Sick, 75–79 | $h_{2,75-79,5}$ | 0.173391 | 0.173391 | 0.173393 | 0.173392 | 0.173391 | 767,540 | 679,591 |
| Sick, 80–84 | $h_{2,80-84,5}$ | 0.128352 | 0.128352 | 0.128352 | 0.128354 | 0.128353 | 531,006 | 471,380 |
| Sick, 85–89 | $h_{2,85-89,5}$ | 0.090954 | 0.090954 | 0.090951 | 0.090953 | 0.090952 | 302,549 | 263,597 |
| Sick, 90–94 | $h_{2,90-94,5}$ | 0.062018 | 0.062017 | 0.062017 | 0.062018 | 0.062016 | 142,886 | 116,784 |
| Sick, 95–99 | $h_{2,95-99,5}$ | 0.040748 | 0.040748 | 0.040748 | 0.040749 | 0.040749 | 50,778 | 41,546 |
| Sick, 100–104 | $h_{2,100-104,5}$ | 0.025662 | 0.025664 | 0.025662 | 0.025664 | 0.025663 | 7,555 | 10,895 |
| Sick, 105–109 | $h_{2,105-109,5}$ | 0.015212 | 0.015214 | 0.015216 | 0.015213 | 0.015214 | 1,048 | 1,187 |
| Sick, 110–114 | $h_{2,110-114,5}$ | 0.008089 | 0.008087 | 0.008090 | 0.008084 | 0.008092 | 102 | 121 |
| Sick, 115+ | $h_{2,115+,5}$ | 0.003273 | 0.003272 | 0.003266 | 0.003271 | 0.003274 | 28 | 11 |
| All | $c_{t=5}$ | 1.567749 | 1.567748 | 1.567749 | 1.567749 | 1.567748 | 125,410,713 | 124,012,454 |

NOTE: For display purposes, the population numbers were rounded to the nearest integer using the R function *round*.
[a]Best of the 5 runs.
[b]This number is about 2.88.
[c]This number is about 1.12.

Table C.3: Information from five genetic algorithm runs for time period 10 (2055–2059)

| For state | Associated argument (of form $h_{g,a,t}$ or $c_t$) | Argument value (per state member) in 2005 per capita income units | | | | | Population for present period | Implied population for next period from best run |
|---|---|---|---|---|---|---|---|---|
| | | Run 1 | Run 2 | Run 3 | Run 4[a] | Run 5 | | |
| Healthy, ages 0–4 | $h_{1,0-4,10}$ | 0.210452 | 0.210451 | 0.21045 | 0.210451 | 0.210451 | 10,000,000 | 10,000,000 |
| Healthy, 5–9 | $h_{1,5-9,10}$ | 0.235341 | 0.235341 | 0.235341 | 0.235341 | 0.235341 | 10,762,842 | 10,781,163 |
| Healthy, 10–14 | $h_{1,10-14,10}$ | 0.263151 | 0.263153 | 0.263151 | 0.263151 | 0.263152 | 10,601,296 | 10,631,538 |
| Healthy, 15–19 | $h_{1,15-19,10}$ | 0.294088 | 0.294088 | 0.294088 | 0.294088 | 0.294088 | 10,380,106 | 10,425,882 |
| Healthy, 20–24 | $h_{1,20-24,10}$ | 0.328283 | 0.328283 | 0.328285 | 0.328283 | 0.328283 | 10,096,488 | 10,161,203 |
| Healthy, 25–29 | $h_{1,25-29,10}$ | 0.365745 | 0.365743 | 0.365743 | 0.365745 | 0.365741 | 9,736,660 | 9,823,975 |
| Healthy, 30–34 | $h_{1,30-34,10}$ | 0.406257 | 0.406257 | 0.406257 | 0.406256 | 0.406257 | 9,284,995 | 9,398,601 |
| Healthy, 35–39 | $h_{1,35-39,10}$ | 0.449277 | 0.449277 | 0.449277 | 0.449277 | 0.449277 | 8,724,958 | 8,868,136 |
| Healthy, 40–44 | $h_{1,40-44,10}$ | 0.493749 | 0.493747 | 0.493749 | 0.493749 | 0.493749 | 8,040,888 | 8,215,795 |
| Healthy, 45–49 | $h_{1,45-49,10}$ | 0.537896 | 0.537906 | 0.537900 | 0.537900 | 0.537900 | 7,221,318 | 7,427,884 |
| Healthy, 50–54 | $h_{1,50-54,10}$ | 0.579000 | 0.578999 | 0.578999 | 0.578999 | 0.578998 | 6,264,420 | 6,498,932 |
| Healthy, 55–59 | $h_{1,55-59,10}$ | 0.613155 | 0.613156 | 0.613158 | 0.613157 | 0.613153 | 5,282,057 | 5,439,509 |
| Healthy, 60–64 | $h_{1,60-64,10}$ | 0.635304 | 0.635309 | 0.635302 | 0.635304 | 0.635304 | 4,211,059 | 4,365,704 |
| Healthy, 65–69 | $h_{1,65-69,10}$ | 0.639608 | 0.639607 | 0.639604 | 0.639607 | 0.639608 | 3,110,066 | 3,251,551 |
| Healthy, 70–74 | $h_{1,70-74,10}$ | 0.620577 | 0.620565 | 0.620580 | 0.620578 | 0.620577 | 2,071,023 | 2,187,555 |
| Healthy, 75–79 | $h_{1,75-79,10}$ | 0.575015 | 0.575015 | 0.575016 | 0.575015 | 0.575016 | 1,232,911 | 1,284,299 |
| Healthy, 80–84 | $h_{1,80-84,10}$ | 0.504276 | 0.504265 | 0.504270 | 0.504268 | 0.504279 | 608,354 | 647,673 |
| Healthy, 85–89 | $h_{1,85-89,10}$ | 0.415418 | 0.415422 | 0.415419 | 0.415423 | 0.415423 | 238,347 | 258,772 |
| Healthy, 90–94 | $h_{1,90-94,10}$ | 0.319903 | 0.319898 | 0.319905 | 0.319901 | 0.319903 | 71,753 | 78,396 |
| Healthy, 95–99 | $h_{1,95-99,10}$ | 0.229626 | 0.229625 | 0.229626 | 0.229620 | 0.229632 | 16,086 | 17,517 |
| Healthy, 100–104 | $h_{1,100-104,10}$ | 0.153024 | 0.153031 | 0.153019 | 0.153027 | 0.153018 | 2,655 | 2,831 |
| Healthy, 105–109 | $h_{1,105-109,10}$ | 0.093448 | 0.093446 | 0.093449 | 0.093442 | 0.093466 | 314 | 333 |
| Healthy, 110–114 | $h_{1,110-114,10}$ | 0.050128 | 0.050122 | 0.050098 | 0.050132 | 0.050120 | 28 | 28 |
| Healthy, 115+ | $h_{1,115+,10}$ | 0.020157 | 0.020187 | 0.020110 | 0.020155 | 0.020120 | 2[b] | 2[c] |
| Sick, 0–4 | $h_{2,0-4,10}$ | 0.349753 | 0.349755 | 0.349755 | 0.349753 | 0.349757 | 1,000,000 | 1,000,000 |
| Sick, 5–9 | $h_{2,5-9,10}$ | 0.391499 | 0.391495 | 0.391489 | 0.391494 | 0.391498 | 117,878 | 108,822 |
| Sick, 10–14 | $h_{2,10-14,10}$ | 0.437603 | 0.437609 | 0.437610 | 0.437612 | 0.437618 | 133,823 | 123,860 |
| Sick, 15–19 | $h_{2,15-19,10}$ | 0.488011 | 0.488011 | 0.488021 | 0.488016 | 0.488011 | 165,170 | 153,471 |
| Sick, 20–24 | $h_{2,20-24,10}$ | 0.542272 | 0.542270 | 0.542272 | 0.542283 | 0.542273 | 203,850 | 190,220 |
| Sick, 25–29 | $h_{2,25-29,10}$ | 0.599414 | 0.599406 | 0.599407 | 0.599410 | 0.599409 | 250,942 | 235,290 |
| Sick, 30–34 | $h_{2,30-34,10}$ | 0.657614 | 0.657620 | 0.657611 | 0.657619 | 0.657632 | 307,480 | 289,879 |
| Sick, 35–39 | $h_{2,35-39,10}$ | 0.713969 | 0.713975 | 0.713975 | 0.713977 | 0.713972 | 373,939 | 354,764 |
| Sick, 40–44 | $h_{2,40-44,10}$ | 0.764073 | 0.764072 | 0.764069 | 0.764071 | 0.764078 | 449,575 | 429,685 |
| Sick, 45–49 | $h_{2,45-49,10}$ | 0.801852 | 0.801859 | 0.801858 | 0.801857 | 0.801861 | 531,374 | 512,354 |
| Sick, 50–54 | $h_{2,50-54,10}$ | 0.819863 | 0.819871 | 0.819875 | 0.819879 | 0.819872 | 612,571 | 596,972 |
| Sick, 55–59 | $h_{2,55-59,10}$ | 0.810296 | 0.810303 | 0.810293 | 0.810296 | 0.810288 | 693,635 | 672,441 |
| Sick, 60–64 | $h_{2,60-64,10}$ | 0.767216 | 0.767214 | 0.767219 | 0.767215 | 0.767218 | 750,907 | 734,501 |
| Sick, 65–69 | $h_{2,65-69,10}$ | 0.689791 | 0.689783 | 0.689791 | 0.689794 | 0.689793 | 761,578 | 753,086 |
| Sick, 70–74 | $h_{2,70-74,10}$ | 0.584770 | 0.584771 | 0.584773 | 0.584771 | 0.584772 | 704,026 | 705,278 |
| Sick, 75–79 | $h_{2,75-79,10}$ | 0.465963 | 0.465962 | 0.465968 | 0.465961 | 0.465962 | 587,633 | 582,330 |
| Sick, 80–84 | $h_{2,80-84,10}$ | 0.349613 | 0.349617 | 0.349617 | 0.349620 | 0.349616 | 410,140 | 416,793 |
| Sick, 85–89 | $h_{2,85-89,10}$ | 0.248429 | 0.248433 | 0.248434 | 0.248434 | 0.248435 | 229,101 | 238,281 |
| Sick, 90–94 | $h_{2,90-94,10}$ | 0.168315 | 0.168316 | 0.168315 | 0.168314 | 0.168316 | 99,058 | 104,076 |
| Sick, 95–99 | $h_{2,95-99,10}$ | 0.109146 | 0.109144 | 0.109146 | 0.109147 | 0.109149 | 32,104 | 33,757 |
| Sick, 100–104 | $h_{2,100-104,10}$ | 0.067541 | 0.067545 | 0.067543 | 0.067543 | 0.067544 | 7,696 | 7,959 |
| Sick, 105–109 | $h_{2,105-109,10}$ | 0.039246 | 0.039243 | 0.039243 | 0.039250 | 0.039248 | 1,324 | 1,366 |
| Sick, 110–114 | $h_{2,110-114,10}$ | 0.020424 | 0.020431 | 0.020420 | 0.020430 | 0.020420 | 168 | 168 |
| Sick, 115+ | $h_{2,115+,10}$ | 0.008082 | 0.008077 | 0.008077 | 0.008084 | 0.008081 | 18 | 17 |
| All | $c_t$ | 2.721844 | 2.721844 | 2.721844 | 2.721844 | 2.721844 | 126,382,619 | 128,012,648 |

NOTE: For display purposes, the population numbers were rounded to the nearest integer using the R function *round*.
[a]Best of the 5 runs.
[b]This number is about 2.08.
[c]This number is about 1.94.

Table C.4: Information from five genetic algorithm runs for time period 15 (2080–2084)

| For state | Associated argument (of form $h_{g,a,t}$ or $c_t$) | Argument value (per state member) in 2005 per capita income units | | | | | Population for present period | Implied population for next period from best run |
|---|---|---|---|---|---|---|---|---|
| | | Run 1 | Run 2 | Run 3 | Run 4 | Run 5[a] | | |
| Healthy, ages 0–4 | $h_{1,0-4,15}$ | 0.381452 | 0.381452 | 0.381453 | 0.381452 | 0.381452 | 10,000,000 | 10,000,000 |
| Healthy, 5–9 | $h_{1,5-9,15}$ | 0.430857 | 0.430861 | 0.430857 | 0.430857 | 0.430853 | 10,841,186 | 10,853,381 |
| Healthy, 10–14 | $h_{1,10-14,15}$ | 0.486847 | 0.486847 | 0.486847 | 0.486848 | 0.486847 | 10,730,943 | 10,751,213 |
| Healthy, 15–19 | $h_{1,15-19,15}$ | 0.550118 | 0.550116 | 0.550116 | 0.550116 | 0.550116 | 10,577,278 | 10,608,346 |
| Healthy, 20–24 | $h_{1,20-24,15}$ | 0.621297 | 0.621297 | 0.621297 | 0.621297 | 0.621297 | 10,376,774 | 10,421,327 |
| Healthy, 25–29 | $h_{1,25-29,15}$ | 0.700868 | 0.700868 | 0.700867 | 0.700868 | 0.700868 | 10,117,289 | 10,178,415 |
| Healthy, 30–34 | $h_{1,30-34,15}$ | 0.788993 | 0.788993 | 0.788993 | 0.788993 | 0.788993 | 9,784,085 | 9,865,211 |
| Healthy, 35–39 | $h_{1,35-39,15}$ | 0.885285 | 0.885292 | 0.885285 | 0.885282 | 0.885285 | 9,359,888 | 9,464,599 |
| Healthy, 40–44 | $h_{1,40-44,15}$ | 0.988448 | 0.988448 | 0.988453 | 0.988448 | 0.988450 | 8,825,423 | 8,957,083 |
| Healthy, 45–49 | $h_{1,45-49,15}$ | 1.095776 | 1.095776 | 1.095776 | 1.095776 | 1.095776 | 8,160,982 | 8,322,059 |
| Healthy, 50–54 | $h_{1,50-54,15}$ | 1.202502 | 1.202501 | 1.202502 | 1.202502 | 1.202502 | 7,349,824 | 7,540,790 |
| Healthy, 55–59 | $h_{1,55-59,15}$ | 1.301057 | 1.301056 | 1.301057 | 1.301057 | 1.301057 | 6,384,468 | 6,602,205 |
| Healthy, 60–64 | $h_{1,60-64,15}$ | 1.380464 | 1.380480 | 1.380480 | 1.380470 | 1.380480 | 5,276,842 | 5,512,674 |
| Healthy, 65–69 | $h_{1,65-69,15}$ | 1.426490 | 1.426490 | 1.426490 | 1.426490 | 1.426490 | 4,071,581 | 4,309,724 |
| Healthy, 70–74 | $h_{1,70-74,15}$ | 1.423091 | 1.423090 | 1.423093 | 1.423091 | 1.423091 | 2,857,083 | 3,075,254 |
| Healthy, 75–79 | $h_{1,75-79,15}$ | 1.356645 | 1.356682 | 1.356644 | 1.356647 | 1.356647 | 1,760,735 | 1,935,383 |
| Healthy, 80–84 | $h_{1,80-84,15}$ | 1.222343 | 1.222349 | 1.222356 | 1.222340 | 1.222353 | 928,627 | 1,028,128 |
| Healthy, 85–89 | $h_{1,85-89,15}$ | 1.030343 | 1.030353 | 1.030351 | 1.030355 | 1.030371 | 393,552 | 445,638 |
| Healthy, 90–94 | $h_{1,90-94,15}$ | 0.806340 | 0.806332 | 0.806330 | 0.806333 | 0.806325 | 127,058 | 147,320 |
| Healthy, 95–99 | $h_{1,95-99,15}$ | 0.583167 | 0.583157 | 0.583159 | 0.583173 | 0.583171 | 29,754 | 35,280 |
| Healthy, 100–104 | $h_{1,100-104,15}$ | 0.388173 | 0.388191 | 0.388165 | 0.388191 | 0.388166 | 5,003 | 5,886 |
| Healthy, 105–109 | $h_{1,105-109,15}$ | 0.235055 | 0.235014 | 0.235053 | 0.235054 | 0.235044 | 575 | 689 |
| Healthy, 110–114 | $h_{1,110-114,15}$ | 0.124384 | 0.124342 | 0.124293 | 0.124358 | 0.124296 | 46 | 55 |
| Healthy, 115+ | $h_{1,115+,15}$ | 0.049145 | 0.049140 | 0.049199 | 0.049083 | 0.049018 | 3[b] | 3[c] |
| Sick, 0–4 | $h_{2,0-4,15}$ | 0.637326 | 0.637326 | 0.637326 | 0.637328 | 0.637326 | 1,000,000 | 1,000,000 |
| Sick, 5–9 | $h_{2,5-9,15}$ | 0.721540 | 0.721538 | 0.721531 | 0.721547 | 0.721541 | 79,094 | 73,043 |
| Sick, 10–14 | $h_{2,10-14,15}$ | 0.816440 | 0.816411 | 0.816426 | 0.816436 | 0.816419 | 90,945 | 84,200 |
| Sick, 15–19 | $h_{2,15-19,15}$ | 0.922590 | 0.922582 | 0.922597 | 0.922581 | 0.922586 | 114,349 | 106,233 |
| Sick, 20–24 | $h_{2,20-24,15}$ | 1.040099 | 1.040117 | 1.040122 | 1.040116 | 1.040105 | 143,995 | 134,273 |
| Sick, 25–29 | $h_{2,25-29,15}$ | 1.168260 | 1.168246 | 1.168249 | 1.168248 | 1.168250 | 181,283 | 169,736 |
| Sick, 30–34 | $h_{2,30-34,15}$ | 1.304855 | 1.304869 | 1.304841 | 1.304857 | 1.304854 | 227,832 | 214,296 |
| Sick, 35–39 | $h_{2,35-39,15}$ | 1.445595 | 1.445584 | 1.445570 | 1.445584 | 1.445580 | 285,250 | 269,695 |
| Sick, 40–44 | $h_{2,40-44,15}$ | 1.583002 | 1.582994 | 1.582952 | 1.582955 | 1.582950 | 354,752 | 337,408 |
| Sick, 45–49 | $h_{2,45-49,15}$ | 1.705382 | 1.705386 | 1.705381 | 1.705355 | 1.705377 | 436,425 | 417,990 |
| Sick, 50–54 | $h_{2,50-54,15}$ | 1.796412 | 1.796419 | 1.796412 | 1.796404 | 1.796427 | 527,939 | 509,853 |
| Sick, 55–59 | $h_{2,55-59,15}$ | 1.835628 | 1.835712 | 1.835688 | 1.835674 | 1.835686 | 622,543 | 607,303 |
| Sick, 60–64 | $h_{2,60-64,15}$ | 1.802235 | 1.802225 | 1.802233 | 1.802247 | 1.802233 | 706,560 | 697,879 |
| Sick, 65–69 | $h_{2,65-69,15}$ | 1.681959 | 1.681938 | 1.681944 | 1.681942 | 1.681930 | 757,748 | 760,117 |
| Sick, 70–74 | $h_{2,70-74,15}$ | 1.476995 | 1.476981 | 1.477004 | 1.476977 | 1.476993 | 748,083 | 765,071 |
| Sick, 75–79 | $h_{2,75-79,15}$ | 1.211721 | 1.211715 | 1.211725 | 1.211716 | 1.211724 | 656,133 | 687,250 |
| Sick, 80–84 | $h_{2,80-84,15}$ | 0.927105 | 0.927095 | 0.927118 | 0.927108 | 0.927090 | 497,641 | 526,706 |
| Sick, 85–89 | $h_{2,85-89,15}$ | 0.664281 | 0.664287 | 0.664288 | 0.664276 | 0.664284 | 306,088 | 332,497 |
| Sick, 90–94 | $h_{2,90-94,15}$ | 0.449077 | 0.449068 | 0.449073 | 0.449073 | 0.449070 | 144,631 | 161,470 |
| Sick, 95–99 | $h_{2,95-99,15}$ | 0.288210 | 0.288210 | 0.288210 | 0.288213 | 0.288211 | 49,960 | 57,264 |
| Sick, 100–104 | $h_{2,100-104,15}$ | 0.175547 | 0.175545 | 0.175546 | 0.175538 | 0.175540 | 12,472 | 14,246 |
| Sick, 105–109 | $h_{2,105-109,15}$ | 0.100062 | 0.100064 | 0.100061 | 0.100067 | 0.100071 | 2,133 | 2,497 |
| Sick, 110–114 | $h_{2,110-114,15}$ | 0.051015 | 0.051008 | 0.050987 | 0.051010 | 0.051017 | 251 | 298 |
| Sick, 115+ | $h_{2,115+,15}$ | 0.019773 | 0.019767 | 0.019759 | 0.019749 | 0.019781 | 22 | 26 |
| All | $c_t$ | 4.694764 | 4.694762 | 4.694763 | 4.694764 | 4.694763 | 135,905,128 | 137,990,018 |

NOTE: For display purposes, the population numbers were rounded to the nearest integer using the R function *round*.
[a]Best of the 5 runs.
[b]This number is about 2.73.
[c]This number is about 3.25.

# BIBLIOGRAPHY


Agresti, A. (2002), *Categorical Data Analysis*, Hoboken, NJ: John Wiley & Sons, Inc., 2nd ed.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: The MIT Press.

Carroll, C. J. (2000), "Requiem for the Representative Consumer? Aggregate Implications of Microeconomic Consumption Behavior," in *Papers and Proceedings of the One Hundred Twelfth Annual Meeting of the American Economic Association*, vol. 90, pp. 110–115.

Caselli, F. and Ventura, J. (2000), "A Representative Consumer Theory of Distribution," *The American Economic Review*, 90, 909–926.

Collins, J., Kaufer, D., Vlachos, P., Butler, B., and Ishizaki, S. (2004), "Detecting Collaborations in Text Comparing the Authors' Rhetorical Language Choices in *The Federalist Papers*," *Computers & the Humanities*, 38, 15–36.

Cox, D. R., Fitzpatrick, R., Fletcher, A. E., Gore, S. M., Spiegelhalter, D. J., and Jones, D. R. (1992), "Quality-of-Life Assessment: Can We Keep It Simple?" *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 155, 353–393.

Daniel, B. C. (1993), "Tax Timing and Liquidity Constraints: A Heterogeneous-Agent Model," *Journal of Money, Credit, and Banking*, 25, 176–196.

Greene, W. H. (2003), *Econometric Analysis*, Upper Saddle River, NJ: Pearson Education, Inc., 5th ed.

Hall, R. E. and Jones, C. I. (2004a), "The Value of Life and the Rise in Health Spending," *NBER Working Paper 10737*, 1–42.

— (2004b), "The Value of Life and the Rise in Health Spending," Version 2.0, formerly available at `http://elsa.berkeley.edu/~chad/hx200.pdf`, 1–43.

— (2007), "The Value of Life and the Rise in Health Spending," *The Quarterly Journal of Economics*, 122, 39–72.

Hamada, M., Martz, H. F., Reese, C. S., and Wilson, A. G. (2001), "Finding Near-Optimal Bayesian Experimental Designs via Genetic Algorithms," *The American Statistician*, 55, 175–181.

Hartley, J. H. (1996), "Retrospectives: The Origins of the Representative Agent," *The Journal of Economic Perspectives*, 10, 169–177.

Haupt, R. L. and Haupt, S. E. (2004), *Practical Genetic Algorithms*, Hoboken, NJ: John Wiley & Sons, Inc., 2nd ed.

Hildenbrand, W. and Kneip, A. (2005), "On Behavioral Heterogeneity," *Economic Theory*, 25, 155–169.

Kirman, A. P. (1992), "Whom or What Does the Representative Individual Represent?" *The Journal of Economic Perspectives*, 6, 117–136.

Manton, K. G. and Land, K. C. (2000), "Multidimensional Disability/Mortality Trajectories at Ages 65 and Older: The Impact of State Dependence," *Social Indicators Research*, 51, 193–221.

Manton, K. G., Woodbury, M. A., and Tolley, H. D. (1994), *Statistical Applications Using Fuzzy Sets*, New York: John Wiley & Sons, Inc.

Martel, R. J. (1996), "Heterogeneity, Aggregation, and a Meaningful Macroeconomics," in *Beyond Microfoundations: Post Walrasian Macroeconomics*, ed. Colander, D. C., Cambridge: Cambridge University Press, pp. 127–144.

Meltzer, D. (2003), "Can Medical Cost-Effectiveness Analysis Identify the Value of Research?" in *Measuring the Gains from Medical Research*, eds. Murphy, K. M. and Topel, R. H., The University of Chicago Press, pp. 206–247.

Nicholson, W. (2002), *Microeconomic Theory: Basic Principles and Extensions*, Thomson Learning, Inc., eighth ed.

Nordhaus, W. D. (2003), "The Health of Nations: The Contribution of Improved Health to Living Standards," in *Measuring the Gains from Medical Research*, eds. Murphy, K. M. and Topel, R. H., The University of Chicago Press, pp. 9–40.

Rencher, A. C. (1998), *Multivariate Statistical Inference and Applications*, New York: John Wiley & Sons, Inc.

— (2002), *Methods of Multivariate Analysis*, New York: John Wiley & Sons, Inc., 2nd ed.

Ross, S. M. (2003), *Introduction to Probability Models*, San Diego, CA: Academic Press, eighth ed.

Russell, T. (1995), "Aggregation, Heterogeneity, and the Coase Invariance Theorem," *Japan and the World Economy*, 7, 105–111.

Sakawa, M. (2002), *Genetic Algorithms and Fuzzy Multiobjective Optimization*, Norwell, MA: Kluwer Academic Publishers.

Sakawa, M. and Yauchi, K. (1998), "Coevolutionary Genetic Algorithms for Nonconvex Nonlinear Programming Problems: Revised GENOCOP III," *Cybernetics and Systems*, 29, 885–899.