WHAT ROLE DOES IDENTITY PLAY IN THE PREFERENCE FOR INCOME REDISTRIBUTION?

LOUISE C. KEELY AND CHIH MING TAN

Abstract. Does identity play a role in determining an individual's preferred level of income redistribution? Identity takes on many dimensions - which are salient? Using the data from the General Social Survey, we provide a stylized fact that preferences over redistribution in the United States vary across specific identity groupings related to race, gender, and class. Our empirical results provide guidance for evaluating theoretical treatments of identity and economic decision-making. We find that, although identity may play a role in determining preference structures, that does not fully explain our results. Our results are more consistent with a theory in which identity provides information to agents.

1. Introduction

What determines an individual's preferred level of income redistribution? We present new evidence from the General Social Survey showing that views on income redistribution in the United States differ along racial, gender and class lines. There is evident salience of identity groups for determining individual tolerance for the reduction of income inequality.

Theoretical treatments of how identity factors into preferred income redistribution are of two types. In the first, individual views on redistribution are influenced

Date: November 2003; Preliminary and Incomplete; Comments welcome.

We are grateful for funding under the Robock Award in Empirical Economics from the University of Wisconsin. Hugh Chipman and Wei-Yin Loh provided invaluable advice on the implementation of the tree regression methods used in this paper. Keely thanks WARF for generous research support. Tan thanks the generous support provided by the Program of Fellowships for Junior Scholars, MacArthur Research Network on Social Interactions and Economic Inequality. We appreciate the diligent research assistance of Dan Wei and Zhiguo Xiao. We thank Steven Durlauf for related discussions and Buz Brock, Larry Samuelson, and participants in the UW Economic department's Theory Lunch for comments.

by factors outside of the economic system. Alesina, Glaeser and Sacerdote [3], for instance, posit that individual utility is dependent on the utilities of members of other ethnic groups. They conclude that this awareness of ethnic heterogeneity, or "racism", could be responsible for the divergence in views on inequality across groups. This idea is a special case of the more general framework elucidated by Akerlof and Kranton [2]. That is, identity matters because people care, in an exogenous fashion, about the actions or outcomes of others in the same or different identity groups.[1] This view purports that the relevance of identity to economic decision-making should be described via modifications to the preference structure.

In the second, identity is salient as a source of information about one's outcome in an environment with uncertainty. Members of identity groups share meaningfully similar characteristics in a way made precise in the text. The outcomes of others in an agent's identity group are used to make predictions about her unknown quantity of interest. These ideas are discussed in Manski [25]. In what Manski terms 'expectations interactions', agents extract information from observing the actions and outcomes of others. In this case, identity matters because the actions and outcomes of others are informational inputs into each individual's decision-making, and identity provides a guide to what information is most salient for this process. So, in contrast to the first view, the manner in which other agents influence the original is through the effect their decisions have on the latter's information set. The preference structures of agents per se are taken to be mutually independent.

The contributions of this paper are as follows. We provide a stylized fact that preferences over redistribution in the United States vary across specific identity groupings related to race, gender, and class. To our knowledge, this result is novel. Our empirical results provide guidance for assessing the relative importance of the two types of theoretical treatments of identity described above. We conclude that the evidence weighs in favor of the information-based theories.

In Section 2, the empirical methodologies, Bayesian and GUIDE tree regression analysis, are formalized and the reasons for their use are explained. In Section 3, the data is briefly described and the empirical results are presented. In Section 4, we interpret our results with the use of a theoretical model. Section 5 concludes.

[1] See also Akerlof [1]. He shows that a preference for behaving like those around one's self can lead to equilibrium differences in behavior across a population.

2. Empirical Methodology

2.1. **Tree Regression Analysis.** It is clear that there is substantial heterogeneity across individuals in their preferences for redistribution. We posit that it is possible to group individuals such that preferences within groups are alike, in a way we make precise shortly. Further, we hypothesize that individuals are grouped according to aspects of identity. We determine which identity characteristics, if any, are important for such classification. This is a way of discriminating between different theories put forth regarding the relationship between identity and preferences.

Formally, let $y \in Y$ denote preferences for redistribution and let $x \in X$ be a vector of identity markers. We view the population of individuals as being classified by their identity markers into an unknown number b of subpopulations indexed by j. This classification divides the space of X into b partitions, $\{A_1, ..., A_b\}$. The partitions A_j are mutually exclusive and their union is X :

$$A_j \cap A_k = \emptyset \ \forall j \neq k$$

and

$$\bigcup_{j=1}^{b} A_j = X.$$

For example, suppose $x = (Race, Sex)$ where $Race$ takes on values $\{B, W\}$ and Sex takes on values $\{M, F\}$. Then, a possible set of partitions, $\{A_1, A_2, A_3\}$, is $\{(BF, BM), (WM), (WF)\}$.

The preferences of all members of a given subpopulation are assumed to be drawn from the same distribution. That is, the distribution of y across the whole population, g_y, is a mixture distribution with b components. Each component is a parametric distribution indexed by j, $f_y(\theta_j)$ where θ_j are the distribution parameters for $j = 1, ..., b$. Each subpopulation weight $\pi(A_j)$ for $j = 1, ..., b$ is the proportion of individuals whose identity markers fall into the partition A_j :

$$g_y = \sum_{j=1}^{b} \pi(A_j) f_y(\theta_j)$$

and

$$\sum_{j=1}^{b} \pi(A_j) = 1.$$

The econometric technique we will use to examine the relationship between identity markers and preference for redistribution is tree regression modeling. Tree regression techniques deliver estimates for $\left(b, \{A_j\}_{j=1}^{b}, \Theta = \{\theta_j\}_{j=1}^{b}\right)$. As a by-product, automatic variable selection is achieved in the sense that only the salient combinations of identity markers are selected. Partitions which are not salient are not considered in determining $\{A_j\}_{j=1}^{b}$. That is, all other proposed, nonsalient identity markers are dropped.

2.2. **Bayesian tree regressions.** The first tree regression approach we exploit is Bayesian tree regression modeling. Key references for this are Chipman, George, and McCulloch [13][14].[2] Bayesian tree regression techniques take a distributional approach to the mixture problem. They directly model the distribution of preferences for redistribution conditional on the tree structure by making specific functional assumptions. Specifically, they assume $f_y(\theta_j)$ is a normal or logit distribution.

In the Bayesian context, tree regressions parcel the population into subgroups, each of which has elements which are exchangeable within the groupings.[3] Exchangability makes specific our notion of similarity between individuals within groups. Draper [17] lays out the importance of exchangability in inference.

The Bayesian tree regression procedure starts by defininining priors over unknown quantities, which in this case would consist of the parameters for the component distributions Θ and the structure of the tree $T = \left(b, \{A_j\}_{j=1}^{b}\right)$. Defining each (Θ, T) set as a tree model and using Bayes rule, the posterior probability of each tree model is derived. That is,

$$p(\Theta, T|Y, X) \propto p(Y|X, \Theta, T)\, p(\Theta, T)$$

[2]The CGM Bayesian CART software used in this paper and documentation can be found at the following web address: http://gsbwww.uchicago.edu/fac/robert.mcculloch/research/code/CART/

[3]Intuitively, the notion of exchangeability describes the case when the "labels" or indices of random variables do not hold information about them; that is, for a finite sequence of random variables $X_1, ..., X_k$,

$$p(X_1 = x_1, ..., X_k = x_k) = p(X_{\pi(1)} = x_1, ..., X_{\pi(k)} = x_k)$$

where $\pi(.)$ is a permutation operator.

$$p(\Theta, T) = p(\Theta|T)\, p(T)$$

where the tree prior and priors over parameters are given respectively by $p(T)$ and $p(\Theta|T)$. Because of the above parametric assumptions, it will be possible to obtain an analytical form for the marginal posterior tree distribution, $p(T|Y, X)$ by integrating across the model parameters. Stochastic search methods can then be employed to locate trees with high posterior probability. We refer the reader to the Technical Appendix for details on how priors are selected and on the algorithm used for finding the optimal tree model.

2.3. **Generalized Unbiased Interaction Detection and Estimation (GUIDE) tree regressions.** The second tree regression method we use is GUIDE tree regressions. Loh [22] is the key reference. GUIDE is an extension of the well-established Classification and Regression Tree (CART) methodology by Breiman, Friedman, Olsen, and Stone [11]. GUIDE's innovation is to minimize the variable selection problem found in CART (see Doyle [16]) using a simple test for linear fit.

GUIDE constructs an overly large tree by iteratively dividing the population into increasingly small subsets. At each stage, the given population is divided into exactly two subgroups. This division is done by locating an identity marker, and split value for that marker, that minimize the classical linear regression sum of squared errors across both groups. The overly large tree is then pruned back using a criterion that maximizes fit and penalizes for complexity. The result of this procedure is that individuals are classified into groups, the members of which are homogenous in the statistical sense. In our context, the expected preference for redistribution is the same for all individuals within a group. The reader is referred to the Technical Appendix for details on the GUIDE procedure.

3. Empirical Analysis

3.1. **Data.** As described in the Introduction, data from the General Social Survey (GSS) is used to examine the association of identities and redistribution preferences in the United States. A variety of topics are covered in the survey, such as political activism, child-rearing, religious beliefs, and women's rights. Demographic variables such as the respondent's age, sex, income bracket, socioeconomic status,

and education level are also collected. The samples are intended to be nationally representative of adults over 18, with weighting of certain groups.[4]

In the tree regression analysis, redistribution preference is treated as the scalar dependent variable, y. The proxy for y is a categorical variable from the GSS that asks about views on governmental redistribution to reduce income differences (EQWLTH). The redistribution question is asked in each wave of the GSS between 1978 and 2000.[5] The question reads as follows:

> *Some people think the government in Washington ought to reduce income differences between the rich and the poor, perhaps by raising the taxes of wealthy family or by giving income assistance to the poor. Others think the government should not concern itself with this income difference between the rich and the poor.*
>
> *Here is a card with a scale from 1 to 7. Think of a score of 1 as meaning that the government ought to reduce the income differences between the rich and the poor, and a score of 7 as meaning that the government should not concern itself with such differences. What score between one and seven comes closest to the way you feel?*

Several identity variables in the GSS are used in each tree regression analysis to constitute the vector x. As each variable is treated as an independent variable, identity questions were chosen that are not choice variables of the respondent, and can thus be considered exogenous. The identity variables are the respondent's age in years (AGE), his gender (SEX), his self-reported race (RACE); the region of the US in which he was living at 16 (REGION), whether the respondent was born in the US (BORN), whether the respondent's parents were born in the US (PARBORN), the respondent's mother's highest educational degree as a proxy of socioeconomic status (MADEG), and what religion in which the respondent was raised (RELIG16). A trend variable (YEAR) is also included. Details of these questions, as relevant, are provided in the Appendix. Some of the variables are somewhat crude proxies for the identity variable we seek to capture. However, due

[4]The oversample of blacks in the 1982 and 1987 waves of the GSS were not included.

[5]There are 15 years in total of EQWLTH data. They are: 1978, 1980, 1983, 1984, 1986, 1987, 1988, 1989, 1990, 1991, 1993, 1994, 1996, 1998, 2000.

to data constraints and in the interests of parsimony, we have chosen these variables as proxies of identifiable characteristics.

Many models on redistributive politics contain results in which one's preferred distribution level will depend on whether one's income is below or above the mean income (see, for instance, Benabou and Ok [7]). Put differently, one's preferred distribution should depend on whether one perceives his income to be above or below the mean income. In the case where individuals do not have complete information about the income distribution, the perception may differ from the actual. Nevertheless, in terms of preferred redistribution, it is the perception that is relevant. More will be said on the theoretical link between preferred redistribution, income and identity below.

In that light, and as a check on our results, the same empirical analysis was run for another question in the GSS as for EQWLTH. This question (FINRELA) asks respondents whether they view their family income to be above or below the average of that of other American families. The question reads as follows:

Compared with American families in general, would you say your family income is far below average (1), below average (2), average (3), above average (4), or far above average (5)?

This question is reported in each wave of the GSS between 1977 and 2000.[6]

Summary tables that document the distribution of responses to EQWLTH and FINRELA under each identity variable and overall are available from the authors.

3.2. **Empirical Results.** Both types of regression tree described in Section 2 were performed using the pooled data for all years in which the relevant dependent variable was asked. As a check on the robustness of the results, all of the trees described below were also constructed using data only from the year with the largest number of responses, 1994.

When Bayesian tree regressions are performed, the dependent variable is treated as continuous. This is a reasonable approximation for the income variable. However, the redistribution preference variable EQWLTH and the relative income variable FINRELA are categorical, with values 1-7 and 1-5 respectively. Therefore, Logit

[6]There are 18 years in total for FINRELA in the data. They are: 1977, 1978, 1980, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1993, 1994, 1996, 1998, 2000.

Bayesian tree regressions were also run. To that end, two additional types of trees were constructed using EQWLTH or FINRELA as the dependent variable. One type uses Logit Bayesian tree regression by constructing a binary variable where 0 corresponds to responses 1-3 above and 1 corresponds to responses 4-7 above for EQWLTH. The second type uses Logit Bayesian tree regression by constructing a binary variable where 0 corresponds to response 1-4 above and 1 corresponds to responses 5-7 above for EQWLTH. For FINRELA, the binary variable is constructed to be 0 for responses 1-3 and 1 for responses 4-5, *or* 0 for responses 1-2 and 1 for responses 3-5.

When GUIDE tree regressions are constructed, there is an option for the software to impute missing values. Values can be missing from the GSS either because of non-response when the question was asked (coded NA) or because the respondent was not asked the question (coded NAP). The tree regressions were run three ways: with missing values excluded, imputing just the NA missing values, and imputing both the NA and NAP missing values.

The tree regressions using the redistribution preference variable (EQWLTH) as the dependent variable are discussed first. A subset of the tree regression results are reported here; the other trees are available from the authors. Figures 1-5 provide diagrams of that subset of regression trees. The results for the non-reported trees are consistent with the results focused on in the following discussion.

The regression trees have the following robust features. The first split is between whites and non-whites, according to the RACE variable.[7] Whites are then split by gender (SEX) and socioeconomic status (MADEG). For the trees using all years of the data, whites whose mothers finished less than high school are split off from other whites. For the other whites, whose mothers completed more than high school, there is a split by gender. For the trees using 1994 data only, there is no split by socioeconomic status.

[7]In some trees, the non-white and non-black group (RACE=3) are grouped with whites. This is not surprising; ideally this question would have had more than 3 possible responses. As it stands, it groups together Asians, Hispanics, and others. This group is too socioeconomically and ethnically diverse to draw conclusions about it. Therefore, we focus on the black-white differences, recognizing the shortcomings of doing so. No better variable is available in the GSS for the years we examined in our regression trees.

It is desirable to determine the significance of these trees. Each terminal node of a tree corresponds to a different regression model. In a tree of size one, the same regression model applies to all groups; you could run a single regression using the entire sample, with EQWLTH as the dependent variable. In a tree of size greater than 1, you would apply a different regression model for each node. Thus, a question of interest is: to what extent is the tree of size 1 more likely to be correct than the best tree of greater than size 1 (as determined by the regression tree analysis)?

There is no such statistic available for the GUIDE tree regressions to the best of our knowledge. However, for all of the Bayesian tree regressions, a log-likelihood ratio test can be used. The results of this are presented in Figure 6. In short, each of the tests cannot reject the hypothesis that the best tree of size greater than 1 is the correct model against the hypothesis that the tree of size 1 is the correct model.

Another issue is that none of the tree structures are exactly alike. Meilă [26] proposes two information theoretic measures of how different two trees are. These measures can be used, for the same data set, across estimation paradigms. The construction and properties of these measures are described in the technical appendix. Intuitively, these measures give an indication of how much information the structure of one tree gives about the structure of another. The comparable trees generated with the dependent variable EQWLTH contain virtually the same information according to these measures, as presented in Figure 28.

In sum, the trees suggest that redistribution preferences have important associations with one's race, sex, and class. In particular, the average redistribution preferences will differ by the robust groups in the trees. These groups are: non-whites, low-status whites, other-status female whites, and other-status male whites. Histograms in Figures 7-10 summarizing the distribution of preferences for redistribution across the four robust groups from the regression trees for 1978-2000.[8] Those histograms suggest important differences in redistribution preferences. Redistribution is preferred from most to least by non-whites, low-status whites, other-status white females, and other-status white males. The redistribution preferences of non-whites are heavily skewed in favor of redistribution. The low-status whites also have

[8]Histograms for other splits by race, gender, and socioeconomic status, for 1994 only and for all years, are available from the authors.

a distribution that is skewed in favor of redistribution, but less so than the non-whites. The main difference between white females and white males (overall in 1994 and other-status for all years) is that the women's distribution is more heavily concentrated in the middle of the preferences and less at the upper end (against redistribution) than the men's distribution.

Tree regressions are also run for the dependent variable FINRELA, with the same independent identity variables. The two main trees, for all years of the GSS survey over 1977-2000, using either GUIDE or Bayesian tree regression methods, are reported in Figures 11-16. Other regression trees, for 1994 and imputing missing values under GUIDE, are available from the authors.

The tree splits are largely consistent with those for EQWLTH. Robust splits are between whites and non-whites, between genders, and between the lowest socioeconomic group (MADEG=0) and the others. However, there is an interesting difference: for some of the trees, including those reported in this paper, religion figures into the nodes. In particular, those raised Jewish or in no religion are split out separately from the other religious groups (Protestant, Catholic, and other) and from each other. This result implies that a different regression model determining how the level of family income is perceived relative to the average family income differs across these religious background groups.

The religion variable does not enter into any of the nodes of the EQWLTH regression trees. This contrast suggests that the process determining views of one's income relative to that of others may differ across religions. However, it does not enter importantly into the process determining views on the optimal level of income redistribution by the government. Religious background, particularly being in a minority group such as those raised with no religion or in the Jewish religion, may affect the perception of one's family income relative to the population average. This is an interesting result, though this paper is not the place for a full investigation of its causes and consequences. There appears to be no important difference in policy preference. Moreover, as will be discussed below, there is little evidence of a difference in terms of *actual* household income across religious backgrounds, conditional on other demographic variables.

It still remains to determine the ordering and distribution of responses to FINRELA across identity groups, and to compare them with those for EQWLTH. The groups for which the histograms are constructed are the same as those for EQWLTH. They are presented in Figures 17-20.[9] A key finding is that the probability mass shifts from the low responses to the high responses for both questions as one moves in order from nonwhites to low status whites, to higher status white females and to higher status males. This result is also found when looking at real income (REALINC) categories across the same set of identity groups, as seen in Figures 21-24.

4. Interpretation

The analysis in Section 3 is atheoretical, and does not immediately give insight into the mechanism by which identity markers would determine preferences for income redistribution. In the Introduction, two classes of theoretical models are distinguished, each of which potentially provides interpretation of the empirical results. The two classes are preference-based models and information-based models of identity. Ideally, the empirical results would rule out one class of model but be consistent with the other.

Of course, there is an identification problem, as in much of the interactions literature. However, it will be concluded that the information-based theories should be favored over preference-based theories. This claim is based on two main arguments: one is that the low-level of tree complexity observed in the results of Section 3 is more consistent with information-based theories. The second is that, viewed in conjunction with the results of Section 3, there is strong empirical evidence from other lines of research that supports an information-based theory over a preference-based theory.

4.1. **A formalized information-based theory.**

Although information-based explanations for the importance of identity have been discussed by Manski [25] and elsewhere, there are few formalizations of such ideas. In this paper's setting, the goal of such a model is to understand how long-run divergence in preferred income

[9] A fuller set of histograms corresponding to sex, race, and socioeconomic status groups is available upon request.

redistribution may be related to identity. In a Theoretical Appendix, such a model is presented. In short, the link is via differences in expected returns to education across identity groups.

The model's basic structure is as follows. The variable used to measure preferred income redistribution in the framework is the preferred tax rate. Tax revenue is used to redistribute income via human capital investment, so income redistribution has a productive effect. Identity can have long run effects on preferred income redistribution and observed outcomes, including income and human capital levels, in a setting where identity is used to extract information about the returns to educational expenditure.[10]

The conceptual motivation for assuming differences in returns to education is as follows. When making education and other human capital investment decisions, utility-maximizing agents have in mind an expected return to that investment. In an environment with homogenous returns and full information, the return is trivially equal across all agents. However, in an environment with potentially heterogenous returns and imperfect information, agents will want to extract information by observing the actions of others (consistent with a particular return to education) and the outcomes of others. Agents will not necessarily try to gather information about everyone's actions; they will choose those with whom they are meaningfully similar (as discussed in Section 2). That is, they will extract information about those whose actions and outcomes provide the best predictor of what one's own action and outcome should and will be. This idea is further developed below in Section 4.2 and in the Theoretical Appendix.

In sum, an information-based interpretation of our results is that identity is used as a way to determine the returns to education. Identity is used as a source of information in an environment with uncertainty. The empirical results suggest that the use of racial, gender, and socioeconomic identity markers are particularly relevant.

4.2. **Comparing two hypotheses about why identity matters: preference-based theories and information-based theories.** Information-based theories

[10]The long-run differences across groups is not a general feature of such models. Indeed, initial level differences generally disappear in a steady state equilibrium.

are based on a hypothesis that differences in preferences for redistribution are due solely to economic, or productive, factors. This view is consistent, for instance, with Benabou [6]. In the model developed in the Theoretical Appendix, identity enters via the human capital production function.

Preference-based theories are based on an alternative hypothesis that differences in redistribution preferences across identity groups are due to differences in preference structures across identity groups. That preferences are driven by racism, as in Alesina, Glaeser and Sacerdote [3], is an example of this type of hypothesis.

The empirical results of Section 3 suggest that the information-based theories provide sufficient explanation; there is no need to invoke preference-based theories to understand the findings. Nevertheless, the argument for information-based theories is strengthened with further consideration.

4.2.1. *Empirical evidence on the returns to human capital across identity groups.* There is strong empirical support that returns to human capital investment differ across gender groups and across whites and blacks in the US, via evidence of unexplained wage gaps. Differences in the median wages of men and women and blacks and whites, in the United States, has been extensively documented.[11] Attention has been paid to understanding to what extent these differences are due to variables such as education and job experience, which are - in principle - choice variables - and to what extent they are due to discrimination. Such discrimination could be of the Becker [5] sort - a taste for discrimination, or of the statistical sort - discrimination due to imperfect information on the part of employers (Arrow [4]). Black-white and gender wage differences. At the risk of over-simplifying, the results of studies on black-white differences are summarized as follows. There are differences in black-white wages for both men and women (Neal and Johnson [20],[28] and Neal [27]). A significant part of these differences between men can be traced back to pre-market variables, such as years of schooling, school quality, and family

[11]Histograms for a variable in the GSS that provides self-reported household income (in real dollars) in Figures 21-24 for the robust identity groups listed in Section 3. This variable gives ranges of incomes, rather than exact amounts. This variable REALINC contains 24 responses. These findings are consistent with 2000 census data summaries. Tables with the requisite data, and a full set of histograms, are available from the authors.

background. Claims about the extent of the gap explained by these variables is subject to some debate (see Darity and Mason [15], Holzer and Neumark [19], Johnson and Neal [20]). Labor force participation differences may also be important, and differences in participation choice across races can induce selection bias into the employed samples. Indeed, potential wages of men and women exhibit larger racial gaps than actual wages; this finding is especially important in considering women, whose racial wage gap is considerably smaller than that of males (Neal [27]).

The male-female gap is attributable largely to differences in labor market participation and occupation category, rather than pre-market variables. (Blau [8], Blau and Kahn [9], O'Neill [29]). There is also some evidence that differences in college major between men and women provide explanation for later wage differences (Brown and Corcoran [12], Joy [21], and Turner and Bowen [32]). Moreover, Brown and Corcoran [12] report the finding that the reward for male-dominated majors is higher for men than women.

In both sets of studies, a significant amount of the gaps remain 'unexplained'. In a typical regression with an individual's wage as the dependent variable, the race and gender variables still are significant, after controlling for all other pre-market and market variables that economists have on hand (Darity and Mason [15], Holzer and Neumark [19]). Moreover, it is recognized that educational level and work experience variables are choice variables. Anticipated returns may vary by gender and race. Thus, the effects of anticipated returns to education may be mixed up in these variables, as they affect behavior. Anticipated returns may also contribute to the 'unexplained' part of the wage gaps. For instance, O'Neill [29] argues that anticipated return may be determined in part by norms about future home responsibilities for women. Discussion by Darity and Mason [15], Turner and Bowen [32], and Blau and Kahn [9] all point to a self-fulfilling prophecy of lower anticipated labor force participation as a reason for the male-female wage gap. There is evidence that the effect of school quality on the number of years of schooling is the main mechanism by which school quality affects the male black-white wage gap (Heckman, Lyons and Todd [18]).

Loury [23] emphasizes that choice variables such as years of schooling, occupational choice, and labor market participation, are a result of a set of social and

cultural factors that may include discrimination, but that affect people of different races and genders differently. Social networks present and inform opportunity and expectations regarding the labor market, including returns to education. Social segregation by race or gender can adversely affect the skills acquired.

Thus, Loury argues, premarket differences between black and white men such as years of schooling cannot be dismissed out-of-hand as voluntary differences. These differences may be the result of perceived differences in the return to a year of education that are generated by, for instance, parental and other role models and exposure to media. Those perceived differences may be ex-post rational in that they are self-fulfilling, and persistent. As a result, initial differences across groups can have long run effects on observed human capital, labor force participation, and income.

Similarly, women arguably live in a society where being financially dependent on a male partner is acceptable, if not desirable, and wehre they typically expect to take on a majority of household chores and childcare with the partnership. This potentially affects their decisions about education and employment. In particular, they may build in either the possibility or the actuality of these circumstances. They may choose lower-paying occupations or invest less in their human capital accumulation through years of schooling, choice of college major, and work experience.

Individual incorporation of social networks and norms that differ across racial and gender groups, in addition to overt demand-side discrimination, can play a role in the determination of observed choices. In our framework, we summarize such incorporation simply via differences across groups in the return to education. Such differences result in long run differences in observed human capital investment and income.

Income differences across religious denominations. There is an empirical literature that examines whether earnings, or returns to human capital, differ across religious denomination, holding other background variables equal. With the exception of the study by Tomes [31], all of these studies suffer from the problem that current religious denomination is used, rather than the denomination of one's upbringing. This difficulty is noted by Sander [30] and Tomes [31]. Tomes suggests that using

current religious denomination as an explanatory variable in an earnings regression is akin to including a dummy for golf club membership. The point is that people may sort into religious denominations by income or education.

This issue is not purely academic. Most studies find that current religious denomination is a significant predictor of earnings. In contrast, Tomes [31] finds, using GSS data, that religious background is not a significant predictor of earnings with other family background variables controlled for. We know of no other studies that examine the issue in a correct way, and thus rely on Tomes's findings to conclude that the evidence does not support the hypothesis that returns to education vary across religious denomination. If preference-based theories were correct, a top candidate for an identity marker in the US that could lead to heterogeneous preferences, but not to different returns to education, is religious background. The fact that religious background is neither an identity marker in the regression trees for EQWLTH nor an important determinant of wage heterogeneity lends support for the information-based theories.

In summary, one interpretation of the empirical results in Section 3 is that differences in returns to education in the robust identity groups leads to the difference in preference for income distribution across groups. This interpretation is an information-based explanation for why identity matters. The explanation is consistent with low-level of observed complexity in the four robust identity groups: non-whites, whites with low socio-economic status, white females with other socioeconomic status, and white males with other socioeconomic status. The explanation, in light of the empirical results of Section 3, is also consistent with the well-documented 'unexplained' wage gap in the US between men and women and blacks and white, and with the absence of such a gap across religious background groupings.

5. Conclusion

In this study, we provide evidence of salient political groupings according to their preference over income redistribution. These groupings correspond to division among three identity groups from a set of eight. The identity groups correspond to race, gender, and socioeconomic status.

The mechanism that relates identity groupings to views on redistribution is not obvious. Divergence of views in the US clearly seems to depend on the multiplicity of identity groupings. It is not possible to say definitively whether these differences are due to economic factors, such as assessment of the returns to education, or due to extra-economic factors such as 'racism'. Nevertheless, it is argued that strong support exists for the hypothesis that identity enters into an individual's decision-making because of expectations interactions. Economic factors can alone account for our empirical findings: preferences for redistribution differ across identity groups because of the use of identity as an information extraction tool.

References

[1] G. A. Akerlof. Social distance and social decisions. *Econometrica*, 65(5):1005–1027, 1997.

[2] G. A. Akerlof and R. E. Kranton. Economics and identity. *Quarterly Journal of Economics*, 115(3):715–753, 2000.

[3] A. Alesina, E. Glaeser, and B. Sacerdote. Why doesn't the United States have a European-style welfare state? *Brookings Papers on Economic Activity*, 2, 2001.

[4] K. J. Arrow. The theory of discrimination. In O. Ashenfelter and A. Rees, editors, *Discrimination in the Labor Markets*, pages 3–33. Princeton University Press, 1973.

[5] G. S. Becker. *The Economics of Discrimination*. Chicago University Press, 1957.

[6] R. Benabou. Inequality and growth. Working Paper 5658, NBER, 1996.

[7] R. Benabou and E. A. Ok. Social mobility and the demand for redistribution: The POUM hypothesis. *The Quarterly Journal of Economics*, 2001.

[8] F. Blau. Trends in the well-being of American women, 1970-1995. *Journal of Economic Literature*, 36(1):112–165, 1998.

[9] F. Blau and L. Kahn. Gender differences in pay. *Journal of Economic Perspectives*, 14(4):75–99, 2000.

[10] R. Breen and C. Garcia-Penalosa. Bayesian learning and gender segregation. *Journal of Labor Economics*, 20(4):899–922, 2002.

[11] L. Breiman, J. Friedman, R. Olsen, and C.J.Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[12] C. Brown and M. Corcoran. Sex-based differences in school content and the male-female wage gap. *Journal of Labor Economics*, 15(3):117–126, 1997.

[13] H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Society*, 93:935–960, 1998.

[14] H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian treed models. *Machine Learning*, 48:299–320, 2002.

[15] W. Darity and P. Mason. Evidence on discrimination in employment: Codes of color, codes of gender. *The Journal of Economic Perspectives*, 12(2):63–90, 1998.

[16] P. Doyle. The use of automatic interaction detector and similar search procedures. *Operational Research Quarterly*, 24:465–467, 1973.

[17] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, series B*, 57(1):45–70, 1995.

[18] J. J. Heckman, T. M. Lyons, and P. E. Todd. Understanding black-white wage differentials, 1960-1990. *American Economic Review*, 90(2):344–349, 2000.

[19] H. Holzer and D. Neumark. Assessing affirmative action. *Journal of Economic Literature*, 38(3):483–568, 2000.

[20] W. Johnson and D. Neal. xx. In C. Jencks and S. E. Mayer, editors, *The Black-White Test Score Gap*. xx, 1998.

[21] L. Joy. Do colleges shortchange women? Gender differences in the transition from college to work. *American Economic Review*, 90(2):471–475, 2000.

[22] W.-Y. Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386, 2002.

[23] G. C. Loury. Discrimination in the post-civil rights era: Beyond market interactions. *Journal of Economic Perspectives*, 12(2):117–126, 1998.

[24] C. F. Manski. Dynamic choice in social settings: Learning from the experiences of others. *Journal of Econometrics*, 58(1-2):121–136, 2000.

[25] C. F. Manski. Economic analysis of social interactions. *Journal of Economic Perspectives*, 14(3):115–136, 2000.

[26] M. Meila. Comparing clusterings. Working paper, University of Washington, October 2002.

[27] D. Neal. The measured black-white wage gap among women is too small. Working paper, University of Chicago, 2003.

[28] D. Neal and W. Johnson. The role of premarket factors in black-white wage differences. *Journal of Political Economy*, 104:869–895, 1996.

[29] J. O'Neill. The gender gap in wages, circa 2000. *American Economic Review*, 93(2):309–314, 2003.

[30] W. Sander. Catholicism and the economics of fertility. *Population Studies*, 46(3):477–489, 1992.

[31] N. Tomes. The effects of religion and denomination on earnings and the returns to human capital. *The Journal of Human Resources*, 19(4):472–488, 1984.

[32] S. Turner and W. Bowen. Choice of major: The changing (unchanging) gender gap. *Industrial and Labor Relations Review*, 52(2):289–313, 1999.

6. Technical Appendix

6.1. **Bayesian tree regression models.** Recall that a Bayesian tree regression model is defined as a parameter-tree set (Θ, T). By Bayes rule, we have that,

$$p(\Theta, T|Y, X) \propto p(Y|X, \Theta, T)\, p(\Theta, T)$$

where the tree prior can be written as,

$$p(\Theta, T) = p(\Theta|T)\, p(T)$$

That is, as a product of terminal node model priors with tree priors. We now describe the form of these priors, how hyperparameters are selected, and finally, how the "best" tree and parameter estimates are arrived at. For further details, we refer the reader to Chipman, George, and McCulloch [13],[14].

6.1.1. *Tree Prior $P(T)$.* In effect, the tree prior is arrived at implicitly through a stochastic tree generation process. Structurally, a (binary) tree consists of nodes which are either terminal, or split into left and right children nodes. At each of these splits, a tree regressor variable has to be decided upon, and some splitting value assigned for the variable so as to define the left and right nodes. Therefore, a tree can be generated using the following algorithm (see Chipman, George, and McCulloch [13]):

(1) Begin by setting T to be the trivial tree consisting of a single root (and terminal) node denoted η.
(2) Split the terminal node η with probability $p_{split}(\eta, T)$.
(3) If the node splits, assign it a splitting rule ρ according to the distribution $p_{rule}(\rho|\eta, T)$, and create the left and right children nodes.
(4) Let T denote the newly created tree, and apply steps 2 and 3 with η equal to the new left and right children.

The splitting probability $p_{split}(\eta, T)$ is modeled as follows,

$$p_{split}(\eta, T) = \alpha (1 + d_\eta)^{-\beta}$$

where d_η is the depth of the node η (i.e., the number of splits above η). Intuitively, if the term in the RH brackets were taken out so that the probability of a node splitting was set to a constant α, then tuning the hyperparameter α would control

the probability of obtaining either larger or smaller size trees (that is, trees with more or fewer terminal nodes). Including the term in the brackets, we see that tuning β essentially penalizes for more complex trees with deep splits. The idea is to penalize overfitting.

The splitting rule $p_{rule}(\rho|\eta, T)$ which assigns the split value (for the chosen predictor) that defines the left and right children nodes is modeled as follows. At every split, a predictor is chosen randomly (uniformly so) from the set of all predictors. If the chosen predictor is ordinal, then the split value is chosen uniformly from the available observed values of the predictor. If the chosen predictor is categorical, then the split value is chosen uniformly from the available categories that define the predictor. Chipman, George, and McCulloch [13] refer to this specification for $p_{rule}(\rho|\eta, T)$ as the *uniform specification of* p_{rule}.

In terms of actually choosing the hyperparameter values, we take Chipman, George, and McCulloch's advice to first generate histograms which give the prior distribution on the number of terminal nodes for trees. That is, the prior distribution over tree sizes. And then to choose the set (α, β) which generates the prior distribution that best reflects our prior preferences for tree sizes. We chose $(\alpha, \beta) = (0.5, 0.5)$ to give a prior distribution that appears as in Figure 25. We chose this particular prior distribution because it is conservative in the following senses. The choice of this prior puts the largest amount of mass on the size 1 tree (simple linear regression model) and then tapers downwards with increasing tree sizes. So, this prior puts less weight on more complex nonlinear regression structures and puts more weight on a simple linear one. In this way, if our estimation results in more than one terminal node, we have not assumed our results and avoid overfitting of the model.

With this set of hyperparameter values, the prior mean size of trees is given to be about 2. This prior reflects an assumption that American society may be fragmented, or polarized, by identity markers, but the fragmentation does not result in large numbers of fundamentally different groups. In the framework of Section 2, this fundamental difference would be described by each group having its own process by which redistribution preferences are determined. This assumption seems

reasonable. The United States is a stable two-party democracy, without civil or military unrest that threatens the political or economic system.

6.1.2. *Model Prior* $P(\Theta|T)$. The specific choice of model priors will depend on the choice of likelihood models. In this paper, we consider two likelihood models, the mean shift normal model, which we then associate with a standard normal-gamma conjugate prior, and the logit model where a prior is specified for the standard deviation of the logit parameter.

The standard normal model with normal-gamma conjugate priors is specified in the following manner,

$$y|\Theta, T, x \sim N\left(\mu, \sigma^2\right)$$

$$\mu|\sigma \sim N\left(\bar{\mu}, \frac{\sigma^2}{a}\right)$$

$$\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}$$

where the hyperparameters $(\nu, \lambda, \bar{\mu}, a)$ are to be chosen. The idea is then to choose values which reflect, as much as possible, prior noninformativeness. Following Chipman, George, and McCulloch [14], we choose $\nu = 3$, which is interpreted as giving prior information about σ equivalent to that which is contained in 3 observations. Letting s be the classical unbiased estimate of σ based on a linear regression fit for the data, we wish to choose λ to reflect the idea that for each terminal node model, the σ associated with these models should be smaller than s but perhaps not too much smaller. One way to do this, is to choose a quantile q such that $\Pr(\sigma < s) = q$, and then to use the implied value of λ since,

$$\lambda = \frac{s^2\Phi_\upsilon^{-1}(1-q)}{\nu}$$

where Φ_υ is the cumulative distribution function for the chi-squared distribution with ν degrees of freedom. To this end, we chose $\lambda = 0.404s^2$ which corresponds to $q = 0.75$. It should be noted that the CGMBayesianCART software initially transforms the data variables into mean 0 and range 1 variables. Hence, noninformativeness would mean that we select $\bar{\mu} = 0$. Finally, to choose a, first, note that the marginal distribution for μ is given by $t_\upsilon\sqrt{\frac{\lambda}{a}}$ where t_υ is the t distribution with ν degrees of freedom. Hence, we may choose a by choosing a c such

that $\Pr(-c < \mu < c) = 0.95$ since the marginal distribution for μ yields $a = \frac{\lambda 3.18^2}{c^2}$. Following Chipman, George, and McCulloch [14], we choose $c = 3$.

The logit model is given by

$$\Pr(Y = 1) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$

In this case, the CGMBayesianCART software allows us to specify the prior standard deviation of the coefficients β. In our context, only an intercept is included and the prior standard deviation of β is chosen equal to 10 as recommended.

6.1.3. *Metropolis-Hasting Algorithm.* Each of the above models yields an analytical form for the marginal likelihood obtained by integrating across the parameter space,

$$p(Y|X,T) = \int p(Y|X,\Theta,T)\, p(\Theta|T)\, d\Theta$$

Combining the above with the tree priors, we get the posterior probability for trees,

$$p(T|X,Y) \propto p(Y|X,T)\, p(T)$$

The idea is now to use a Metropolis-Hasting algorithm to simulate a sequence of trees $T^0, T^1, T^2, \ldots\ldots$ which converge in distribution to the posterior $p(T|X,Y)$. The algorithm is as follows. Start with some initial tree T^0 and simulate the transition from any current tree T^i to T^{i+1} in the following manner:.

(1) Generate a candidate tree T^* with probability distribution $q\left(T^i, T^*\right)$.
(2) Set $T^{i+1} = T^*$ with probability

$$\alpha\left(T^i, T^*\right) = \min\left\{\frac{q\left(T^*, T^i\right)}{q\left(T^i, T^*\right)} \frac{p(Y|X,T^*)\, p(T^*)}{p(Y|X,T^i)\, p(T^i)}, 1\right\}$$

Otherwise, set $T^{i+1} = T^i$.

Under weak conditions, the sequence generated by the above algorithm will be a Markov chain with limiting distribution $p(T|X,Y)$.

The specification for the transition kernel $q\left(T^i, T^*\right)$ is obtained by randomly choosing among four steps:

(1) **GROW** : Randomly pick a terminal node. Split it into two new ones by randomly assigning it a splitting rule according to p_{rule} used in the tree prior.

(2) **PRUNE** : Randomly pick a parent of two terminal nodes and turn it into a terminal node by collapsing the nodes below it.

(3) **CHANGE** : Randomly pick an internal node, and randomly reassign it a splitting rule according to p_{rule} used in the tree prior.

(4) **SWAP** : Randomly pick a parent-child pair which are both internal nodes. Swap their splitting rules unless the other child has an identical rule. In that case, swap the splitting rule of the parent with that of both children.

How do we then go about choosing the good trees generated by this process? One way to do this would be to compare the (unnormalized) posterior probabilities of trees, $p(Y|X,T)\,p(T)$. However, as pointed out by Chipman, George, and McCulloch, there is a subtle problem to using this approach. The problem is that two trees of equal size can have very different prior probabilities, and consequently different posterior probabilities. For example[12], suppose we had a categorical variable x_1 which took values of either 1 or 2, and another categorical variables x_2 which took values $1, 2, 3, ..., 100$. Then, a binary tree that splits on x_1 has prior probability $\frac{1}{2}\text{x}\frac{1}{1}$ since there are two variables to split on, and given that x_1 is chosen, there is only one assignment possibility given the values x_1 takes. If the tree splits on x_2, then a specific tree will have prior probability $\frac{1}{2}\text{x}\frac{1}{99}$ of occurring, since there are 99 unique split values for this variable. This means that the posterior probability of a given tree can be "diluted" by the prior depending on what variables are split on. Comparing individual trees using posterior probabilities are therefore misleading. Using posterior probabilities for comparisons only make sense if we are looking at collections of trees. In this example, for instance, it may be possible that there are a dozen different trees that split on x_2 that are all "good". Each might have small posterior probability, but when taken together, they might have greater posterior probability than a single tree splitting on x_2. The suggestion is therefore to simply use the marginal likelihood $p(Y|X,T)$ instead for locating good trees. The trees reported in this paper are those with the highest such values for runs with $50,000$ iterations per chain and for 200 restarts.

[12]We thank Hugh Chipman for providing us with this example.

6.2. **Generalized Unbiased Interaction Detection and Estimation (GUIDE).** GUIDE is an extension of the established CART software (see Breiman et al (1984)) that minimizes the variable selection problem found in the latter (see Doyle (1973)) as well as making possible more complicated terminal node model structures. It solves the regressor (splitting variable) selection bias problem using a simple test for linear fit to select a splitting variable. The choice of split points for selected splitting variables is carried out by choosing the value in the support of the splitting variable that minimizes the sum of squared residuals across regressions of each group. The overall size of the tree is determined using cost-complexity pruning with V-fold cross-validation; again, similar with the CART procedure. In this appendix, we reproduce and detail the key GUIDE algorithms from the main reference used in this paper. Loh (2002) provides more details about GUIDE as well as a description of other tree generating options within GUIDE.

First, define the following classes of covariate variables:

- $n - variable$: a numerical-valued predictor used to fit the terminal node regression model and to split the nodes in the tree;
- $f - variable$: a numerical-valued predictor used to fit the terminal node regression model but not to split the nodes in the tree;
- $s - variable$: a numerical-valued predictor used to split the nodes in the tree but not to fit the terminal node regression model;
- $c - variable$: a categorical predictor used to split the nodes in the tree but not to fit the terminal node regression model.

The first two algorithms determine the choice of the splitting variable at each node of the tree.

Chi-square tests for linear fit.

- Obtain the residuals from a linear model fitted to the n- and f-variables, leaving out the s- and c-variables.
- For each n-variable, divide the data into four groups at the sample quartile; construct a 2×4 contingency table with the signs of the residuals (positive versus non-positive) as rows and the groups as columns; count the number

of observations in each cell and compute the χ^2-statistic and its theoretical p-value from a χ^2_3 distribution.

- Do the same for each s- and c-variable. For the latter, the categories of the variable form the columns of the table. Columns with zero column totals are omitted.
- To detect interactions between each pair of n-variables (X_i, X_j), divide the (X_i, X_j)-space into four quadrants by splitting the range of each variable into two halves at the sample median; construct a 2×4 contingency table using the residual signs as rows and the quadrants as columns; compute the χ^2-statistic and p-value. Again, columns with zero column totals are omitted.
- Do the same for each pair of s-variables.
- Also do the same for each pair of c-variables using their value pairs to divide the sample space. For example, if X_i and X_j take c_i and c_j unique values, respectively, the χ^2-statistic and p-value are computed from a table with 2 rows and number of columns equal to $c_i c_j$ less the number of zero columns.
- Compute a χ^2-statistic and p-value for each pair (X_i, X_j) where X_i is an n-variable and X_j is a c-variable. If X_j has c categories, the table has 2 rows and number of columns equal to 2c less the number of zero columns.
- Similarly, compute a χ^2-statistic and p-value for each pair (X_i, X_j) where X_i is an s-variable and X_j is a c-variable.
- Finally, do the same for each pair where X_i is an s-variable and X_j is a n-variable as in step 4.

Choosing the splitting variable.

- Note that 9 sets of Chi-square tests are computed: 3 sets to detect curvature in the n-, s-, and c-variables, 3 sets to detect interactions between pairs of variables of the same type, and 3 sets to detect interactions between pairs of predictors of different types.
- If the smallest p-value comes from a curvature test, the associated variable is selected to split the node.

- Suppose instead that a pair of variables is selected because their interaction test is the most significant among the curvature and interaction tests.
- If neither is a n-variable, choose the one with the smaller curvature p-value.
- If both are n-variables, temporarily split the node along the sample mean of each variable; choose the variable whose split yields the smaller total SSE.
- If exactly one is an n-variable, choose the other variable.

Once a splitting variable (call it X_j) has been chosen, we need to determine the split value for that variable. This is done in the next algorithm.

Choosing the split value.

- Consider the two partitions of the sample space $Y \times X$ defined as follows,

$$A_1^j(s) = \{(y_i, X_i) \in Y \times X | x_j^i \leq s\} \tag{6.1}$$

$$A_2^j(s) = \{(y_i, X_i) \in Y \times X | x_j^i > s\} \tag{6.2}$$

  where $i = 1, ..., n$ indexes obervations and $x_j^i \in X_j$ for all i.
- The task is to determine the value s.
- Let $\hat{\beta}^1_{(j,s)}$ be the OLS estimator of the regression of Y on X for the subset of observations that conforms to the partition $A_1^j(s)$, and $\hat{\beta}^2_{(j,s)}$ be the OLS estimator of the regression of Y on X for observations conforming to partition $A_2^j(s)$.
- Find the split value s for splitting variable X_j that minimizes the sum of squared residuals (SSR):

$$\sum_{(y_i, X_i) \in A_1^j(s)} \left(y_i - X_i \hat{\beta}^1_{(j,s)}\right)^2 + \sum_{(y_i, X_i) \in A_2^j(s)} \left(y_i - X_i \hat{\beta}^2_{(j,s)}\right)^2 \tag{6.3}$$

To grow a tree, therefore, GUIDE starts with the set of all observations and applies the three algorithms above to find a splitting variable and a split value leaving two mutually exclusive subsets of observations the union of which forms the set of all observations. It then continues to apply this same procedure to each of the resultant subsets, and then to the subsets of observations resulting from those, and so forth iteratively until the number of observations in the subset falls below a certain predetermined value. In our exercises, we take this minimum number

of observations to be the default value set by GUIDE. After the tree is grown, it has to be "pruned" in order to avoid overfitting the data. This is done using Cost Complexity pruning.

Cost Complexity Pruning

- First denote the fully-grown tree to be pruned as T_0. Now, consider a sub-tree $T \subset T_0$ with b terminal nodes.
- Define the Cost Complexity criterion for given α (as applied to sub-tree T) as,

$$C_\alpha(T) = \sum_{m=1}^{b} \sum_{(y_i, X_i) \in TN(m)} \left(y_i - X_i \hat{\beta}_m \right)^2 + \alpha \cdot b \tag{6.4}$$

- The idea behind the use of the Cost Complexity criterion is to attempt to minimize SSR but to penalize overfitting through the use of an overly complex tree.
- Denote the tree for which $C_\alpha(T)$ is minimized (for given α) as $T_\alpha \subseteq T_0$.
- The SSR for T_α is obtained by cross-validation.
- Choose the tree for which α minimizes the cross-validated SSR.

For the GUIDE generated trees reported in this paper, the covariate in the terminal node model consists only of a constant. V was set to the maximum value available.

6.3. **Tree Comparisons.** The method we use to measure the degree of similarity between the trees is outlined by Meilă [26]. An important advantage to her method over other existing ones is the ability to compare trees generated with different algorithms. Therefore, we can compare trees generated with GUIDE and using Bayesian methods and with the same underlying dataset.

The measures are summarized here. Some notation is necessary. A clustering Σ is a partition of the data set D into sets $C_1, C_2, ..., C_K$ such that

$$C_k \cap C_l = \emptyset \text{ and } \cup_{k=1}^{K} C_k = D.$$

Let the number of data points in D and in cluster C_k be n and n_k respectively. Thus,

$$n = \sum_{k=1}^{K} n_k$$

and it is assumed that $n_k > 0$, so K is the number of non-empty clusters. Let a second clustering of the same data set D be $\Sigma' = \{C'_1, C'_2, ..., C'_{K'}\}$ with cluster sizes n'_k. The two clusterings may have different numbers of clusters.

Define the probability of any element of a data set D being in a cluster, C_k as

$$P(k) = \frac{n_k}{n}.$$

The entropy associated with a clustering Σ is

$$H(\Sigma) = -\sum_{k=1}^{K} P(k) \ln P(k).$$

Entropy gives a measure of the uncertainty associated with the clustering.

Define the variable $P(k, k')$ that represents the probability that a point belongs to C_k in clustering Σ and C'_k in clustering Σ' as

$$P(k, k') = \frac{|C_k \cap C'_k|}{n}.$$

Define $I(\Sigma, \Sigma')$, a measure of the mutual information between clusterings, as

$$I(\Sigma, \Sigma') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \ln \frac{P(k, k')}{P(k) P'(k')}.$$

Intuitively, $I(\Sigma, \Sigma')$ tell us how much information the clustering Σ gives us about the clustering Σ'. The bounds on $I(\Sigma, \Sigma')$ are:

$$0 \leq I(\Sigma, \Sigma') \leq \min[H(\Sigma), H(\Sigma')].$$

Equality in the above expression indicates that one clustering completely determines the other. An example of this determination is when one clustering Σ is formed by merging two or more clusters of the other clustering Σ'.

The comparison measure proposed by Meilă is

$$\begin{aligned} VI(\Sigma, \Sigma') &= H(\Sigma) + H(\Sigma') - 2I(\Sigma, \Sigma') \\ &= [H(\Sigma) - I(\Sigma, \Sigma')] + [H(\Sigma') - I(\Sigma, \Sigma')]. \end{aligned}$$

This expression is a measure of the variation of information between the two clusterings. The first term $[H(\Sigma) - I(\Sigma, \Sigma')]$ measures the information lost and

$[H(\Sigma') - I(\Sigma, \Sigma')]$ the information gained in going from clustering Σ to Σ'. The bounds on $VI(\Sigma, \Sigma')$ are

$$0 \leq VI(\Sigma, \Sigma') \leq \ln n$$

or, if both Σ and Σ' have at most $K^* \leq \sqrt{n}$ clusters each, then

$$0 \leq VI(\Sigma, \Sigma') \leq 2 \ln K^*.$$

The distributional properties of $I(\Sigma, \Sigma')$ and $VI(\Sigma, \Sigma')$ is an open research question in the statistics literature. Therefore, it is only possible to calculate them and the possible range of their values for specific examples, and to evaluate each measure based on the range.

In the language of this paper, each clustering corresponds to a tree, and each cluster to a terminal node.

7. Data Appendix

Here identity variables are detailed where their description in the text is incomplete. For the GSS, those variables are RACE, REGION, BORN, PARBORN, MADEG, and RELIG16.

- RACE: *What race would you consider yourself?* (Recorded verbatim and coded)

Responses: White (1), Black (2), Other (3)

- REGION: *In what state or foreign country were you living when you were 16 years old?* (Coded by region)

Responses: New England (1), Middle Atlantic (2), East North Central (3), West North Central (4), South Atlantic (5), East South Central (6), West South Central (7), Mountain (8), Pacific (9), Foreign (10)

New England = Maine, Vermont, New Hampshire, Connecticut, Rhode Island, Massachusetts

Middle Atlantic = New York, New Jersey, Pennsylvania

East North Central = Wisconsin, Indiana, Ohio, Illinois, Michigan

West North Central = Minnesota, Iowa, Missouri, North Dakota, South Dakota, Missouri, Kansas

South Atlantic = Delaware, Maryland, West Virginia, Virginia, North Carolina, South Carolina, Georgia, Florida, District of Columbia

East South Central = Kentucky, Tennessee, Alabama, Mississippi

West South Central = Arkansas, Oklahoma, Louisiana, Texas

Mountain = Montana, Idaho, Wyoming, Nevada, Utah, Colorado, Arizona, New Mexico

Pacific = Washington, Oregon, California, Alaska, Hawaii

- BORN: *Were you born in this country?*

Responses: Yes (1), No (2); don't know responses were treated as missing values

- PARBORN: *Were both of your parents born in this country?*

Responses: Both born in the US (1), One born in the US (2), Neither born in the US (3); don't know responses were treated as missing values

- MADEG: *Respondent's mother's education* (Recoded by GSS from a set of questions regarding years of schooling and degrees attained)

Responses: Less than high school (0), high school (1), Associate/junior college (2), Bachelor's (3), Graduate (4); don't know or NA responses treated as missing values

- RELIG16: *In what religion were you raised?*

Responses: Protestant (1), Catholic (2), Jewish (3), None (4), Other (5)

8. Theoretical Appendix

In this appendix, a framework is developed that is used to highlight how identity can have long run effects on preferred income redistribution and observed outcomes, including income and human capital levels, in a setting where identity is used to extract information. The variable used to measure preferred income redistribution in this framework is the preferred tax rate. In the model, tax revenue will be used only to redistribute income. Income redistribution will occur via human capital investment, so income redistribution has a productive effect. It can be shown, in notes available from the authors, that our main point holds when income redistribution has no productive effect.

8.1. **Households.** Let the population be normalized to 1 so that each dynasty is indexed by $i \in [0,1]$. This model is dynamic; time is discrete and indexed by t. Each unit of time corresponds to one generation. The population is differentiated along J dimensions, indexed by $j = 1, ..., J$. There is also some initial ($t = 1$) human capital level for each household, that is assumed equal across all.

Each time $t = 0, 1, 2, ...$ generation lives for two periods $(t, t+1)$. There is only one child per generation per dynasty, so population is constant. Each dynasty considers itself small compared to the population, and thus will take average, or aggregate, variable measures as given at each point in time.

In period t, the agent from dynasty i and generation t, receives a bequest b_t^i from her parent. The agent also receives human capital h_{t+1}^i, as determined below. Human capital is determined in part by tax revenue that funds education.

In period $(t+1)$ the agent from dynasty i and generation t supplies her human capital to the labor market, consumes c_{t+1}^i, leaves bequest b_{t+1}^i for her generation $(t+1)$ offspring. She participates in a voting mechanism to choose a tax rate τ_{t+1} that is used to finance her offspring's education. Parents have preferences over the offspring's human capital level.

The Household's utility maximization problem is to choose c_{t+1}^i and b_{t+1}^i in order to maximize:

$$V_t^i = \left[\beta_c \left(c_{t+1}^i\right)^\rho + \beta_b \left(b_{t+1}^i\right)^\rho + \beta_h \left(h_{t+2}^i\right)^\rho\right]^{\frac{1}{\rho}}$$

subject to the budget constraint,

$$c_{t+1}^i + b_{t+1}^i \leq I_{t+1}^i \left(1 - \tau_{t+1}\right) \tag{8.1}$$

$$I_{t+1}^i = w_{t+1} h_{t+1}^i + r_{t+1} b_t^i \tag{8.2}$$

$$h_{t+2}^i = \phi H_{t+1}^\gamma e_{t+1}^\xi \left(h_{t+1}^i\right)^\theta \tag{8.3}$$

where w_{t+1} is the wage rate at time $t+1$, r_{t+1} is the interest rate at time $t+1$, H_{t+1} is the average human capital level at time $t+1$, e_{t+1} is expenditure per child on education. Note that equation (8.3) gives the human capital production function. The agent takes e_{t+1} as given. As will be described below, e_{t+1} will be determined by the tax rate τ_{t+1}, that will in turn be derived from a voting mechanism in which the agent participates.

The first order conditions for the utility maximization problem are:

$$b_{t+1}^i = I_{t+1}^i \left(1 - \tau_{t+1}\right) \Psi_c \tag{8.4}$$

$$c_{t+1}^i = I_{t+1}^i \left(1 - \tau_{t+1}\right) \Psi_b \tag{8.5}$$

where

$$\Psi_b = \left[\frac{\beta_b^{\frac{1}{\rho-1}}}{\beta_c^{\frac{1}{\rho-1}} + \beta_b^{\frac{1}{\rho-1}}} \right] > 0$$

$$\Psi_c = \left[\frac{\beta_c^{\frac{1}{\rho-1}}}{\beta_c^{\frac{1}{\rho-1}} + \beta_b^{\frac{1}{\rho-1}}} \right] > 0.$$

8.2. **Production.** Expressions for the wage rate and interest rates are now solved for. A Cobb-Douglas production function is assumed:

$$\begin{aligned} Y_t &= F\left(K_t, H_t\right) \\ &= AK_t^{\alpha} H_t^{1-\alpha} \end{aligned} \tag{8.6}$$

where K_t is aggregate capital, or,

$$y_t = Ak_t^{\alpha} \tag{8.7}$$

where $k_t = \frac{K_t}{H_t}$ and $y_t = \frac{Y_t}{H_t}$. The market structure is assumed perfectly competitive. Therefore, it is appropriate to solve a representative firm's problem:

$$\{K_t, H_t\} = \arg\max \left(AK_t^{\alpha} H_t^{1-\alpha} - w_t H_t - r_t K_t \right) \tag{8.8}$$

taking wage rate, w_t, and rental rate of capital, r_t, as given. The first order conditions for the representative firm's profit maximization problem are:

$$r_t = \alpha A k_t^{\alpha - 1} \tag{8.9}$$

$$w_t = \left(1 - \alpha\right) A k_t^{\alpha} \tag{8.10}$$

8.3. **Market Clearing Conditions.** Conditions to clear the labor and capital markets and to ensure internal consistency of the model are as follows.

The aggregate level of bequests at time t is given by:

$$\begin{aligned} B_t &= \int_0^1 b_t^i di \\ &= (1-\tau_t)\,\Psi_c \int_0^1 I_t^i di \end{aligned} \tag{8.11}$$

where the second equality is given by equation (8.4).

The aggregate expenditure on education is given by:

$$\begin{aligned} e_{t+1} &= \tau_{t+1} \int_0^1 I_{t+1}^i di \\ &= \left(\frac{\tau_{t+1}}{1-\tau_{t+1}}\right) \Psi_c^{-1} B_{t+1} \end{aligned}$$

where the second equality is derived from (8.11) just above.

The aggregate level of capital at time $t+1$ will equal aggregate time t bequests:

$$K_{t+1} = B_t$$

Average human capital across the population is given by:

$$\begin{aligned} H_{t+1} &= \int_0^1 h_{t+1}^i di \\ &\quad \phi H_t^{\gamma} e_t^{\xi} \int_0^1 \left(h_t^i\right)^{\theta} di \end{aligned}$$

Putting this together, the time $t+1$ per capita capital level is:

$$\begin{aligned} k_{t+1} &= \frac{K_{t+1}}{H_{t+1}} \\ &= \left[\frac{\Psi_c^{\xi}}{\phi}\right] \left(\frac{\tau_t}{1-\tau_t}\right)^{-\xi} B_t^{1-\xi} H_t^{-\gamma} \left[\int_0^1 \left(h_t^i\right)^{\theta} di\right]^{-1} \end{aligned} \tag{8.12}$$

Note that all variables influencing k_{t+1} are predetermined at time $t+1$, when generation t makes its decisions.

Solving for equilibrium levels of r_{t+1} and w_{t+1},

$$r_{t+1} = \left[\frac{\alpha A \Psi_c^{-\xi(1-\alpha)}}{\phi^{\alpha-1}}\right] \left(\frac{\tau_t}{1-\tau_t}\right)^{(1-\alpha)\xi} B_t^{\alpha-1} H_t^{\gamma(1-\alpha)} \left[\int_0^1 \left(h_t^i\right)^{\theta} di\right]^{1-\alpha} \tag{8.13}$$

$$w_{t+1} = \left[\frac{(1-\alpha)\, A \Psi_c^{\xi\alpha}}{\phi^{\alpha}}\right] \left(\frac{\tau_t}{1-\tau_t}\right)^{-\alpha\xi} B_t^{\alpha} H_t^{-\gamma\alpha} \left[\int_0^1 \left(h_t^i\right)^{\theta} di\right]^{-\alpha} \tag{8.14}$$

Furthermore, using the above expression for k_{t+1} (8.12), the firm's first order condition (8.9) and the budget constraint definition of income (8.2), it can be shown that,

$$
\begin{aligned}
(8.15)\quad & I^i_{t+1} \\
= & \left[\frac{(1-\alpha)\,A\Psi_c^{\xi\alpha}}{\phi^{\alpha}}\right]\left(\frac{\tau_t}{1-\tau_t}\right)^{-\alpha\xi} B_t^{\alpha} H_t^{-\gamma\alpha}\left[\int_0^1 \left(h_t^i\right)^{\theta} di\right]^{-\alpha}\cdot h^i_{t+1} \\
& + \left[\frac{\alpha A\Psi_c^{-\xi(1-\alpha)}}{\phi^{\alpha-1}}\right]\left(\frac{\tau_t}{1-\tau_t}\right)^{(1-\alpha)\xi} B_t^{\alpha-1} H_t^{\gamma(1-\alpha)}\left[\int_0^1 \left(h_t^i\right)^{\theta} di\right]^{1-\alpha}\cdot b^i_t.
\end{aligned}
$$

Note that τ_{t+1} does not enter the expression for I^i_{t+1}.

8.4. **Household's Preferred Tax Rate.** The question answered here is: What would each dynasty i's generation t agent choose as her utility maximizing tax rate $\hat{\tau}_{t+1}$ holding her h^i_{t+1} and I^i_{t+1} $\forall i$ as given? To answer this question, the indirect utility function is written as:

$$
V_t^i\left(\tau_{t+1}\right) = \left[\begin{array}{c} \beta_c\left(I^i_{t+1}\right)^{\rho}\left[\Psi_b\right]^{\rho}\left(1-\tau_{t+1}\right)^{\rho} + \beta_b\left(I^i_{t+1}\right)^{\rho}\left[\Psi_c\right]^{\rho}\left(1-\tau_{t+1}\right)^{\rho} \\ +\beta_h\phi^{\rho}H_{t+1}^{\gamma\rho}\left(\int_0^1 I^i_{t+1}di\right)^{\xi\rho}\left(h^i_{t+1}\right)^{\theta\rho}\tau_{t+1}^{\xi\rho} \end{array}\right]^{\frac{1}{\rho}}
$$

This expression is concave in τ_{t+1}. Taking the derivative with respect to τ_{t+1}, the first order condition is:

$$
\begin{aligned}
V_t^{i\prime}\left(\hat{\tau}^i_{t+1}\right) &= \frac{1}{\rho}\left(V_t^i\left(\hat{\tau}_{t+1}\right)\right)^{1-\rho}\left[\begin{array}{c} -\rho\left(I^i_{t+1}\right)^{\rho}\left(\beta_c^{\rho}\left[\Psi_b\right]^{\rho}+\beta_b^{\rho}\left[\Psi_c\right]^{\rho}\right)\left(1-\hat{\tau}^i_{t+1}\right)^{\rho-1} \\ +\xi\rho\left(\beta_h\phi^{\rho}H_{t+1}^{\gamma\rho}\left(\int_0^1 I^i_{t+1}di\right)^{\xi\rho}\left(h^i_{t+1}\right)^{\theta\rho}\right)\left(\hat{\tau}^i_{t+1}\right)^{\xi\rho-1} \end{array}\right] \\
&= 0
\end{aligned}
$$

Rearranging,

$$
\frac{\left(\hat{\tau}^i_{t+1}\right)^{\xi\rho-1}}{\left(1-\hat{\tau}^i_{t+1}\right)^{\rho-1}} = \left[\frac{\left(\beta_c\left[\Psi_b\right]^{\rho}+\beta_b\left[\Psi_c\right]^{\rho}\right)}{\xi\beta_h\phi^{\rho}}\right]\frac{\left(I^i_{t+1}\right)^{\rho}}{\left(\int_0^1 I^i_{t+1}di\right)^{\xi\rho}}H_{t+1}^{-\gamma\rho}\left(h^i_{t+1}\right)^{-\theta\rho}.
$$

Note, suppose $\xi = 1$ and $\theta = 0$, then the interpretation is particularly stark: as agent i's income is greater relative to the average, the lower the tax rate she prefers. Without these assumption, the interpretation is less clear; the relationship between an agent's outcome (given by income and human capital) and her preferred tax rate is ambiguous. More will be said on this below.

However, regardless, the optimal (preferred) tax rate $\hat{\tau}_{t+1}^{i}$ is uniquely determined, and,

$$\hat{\tau}_{t+1}^{i} = \hat{\tau}\left(I_{t+1}^{i}\left(h_{t+1}^{i}\left(h_{t}^{i}\right), b_{t}^{i}\right), h_{t+1}^{i}\left(h_{t}^{i}\right); \beta_{c}, \beta_{b}, \beta_{h}, \alpha, \gamma, \xi, \rho, \theta\right).$$

This is depicted in Figure 26.

8.5. **Voting Mechanism.** The institutional assumption made is that there is a majority voting mechanism to choose the tax rate at time t, τ_t. Households are assumed to vote sincerely. Preferences for the tax rate, τ_t, have been established to be single-peaked. Thus, applying the median voter theorem, it is noted that the median voter's preferred tax rate will be chosen.

It is of some interest to establish from which group the median voter will come. Suppose there are four identity groups in the population, of proportions $\mu\lambda$ (male and black or MB), $\mu(1-\lambda)$ (male and white or MW), $(1-\mu)\lambda$ (female and black or FB), and $(1-\mu)(1-\lambda)$ (female and white or FW). There is no population growth. Each group has a unique preferred tax rate.

Assume the initial endowments, denoted by b^j, $j \in [MB, MW, FB, FW]$ are ordered such that b^{FB} is the lowest and b^{MW} is the highest. Furthermore, assume that $\lambda < \mu < 0.5$. Then, given the actual population proportions in the US population, the FW identity group will include the median voter, so that $\forall t$,

$$\tau_{t+1} = \hat{\tau}\left(I_{t+1}^{WF}\left(h_{t+1}^{WF}\left(h_{t}^{WF}\right), b_{t}^{WF}\right), h_{t+1}^{WF}\left(h_{t}^{WF}\right); \beta_{c}, \beta_{b}, \beta_{h}, \alpha, \gamma, \xi, \rho, \theta\right).$$

8.6. **Dynamics And Steady State Solution.**

8.6.1. *The dynamic system.* First, all of the equations needed to fully describe the dynamic system are listed below. The state variables for this system are $\left(b_{t+1}^{i}, h_{t+1}^{i}\right)_{i\in[0,1]}$.

For $i \in [0,1]$,

$$
\begin{aligned}
(8.16) \qquad & b_{t+1}^i \\
= \; & \left[\frac{(1-\alpha)\, A\Psi_c^{1+\alpha\xi}}{\phi^\alpha}\right](1-\tau_{t+1})\left(\frac{\tau_t}{1-\tau_t}\right)^{-\xi\alpha} \\
& \cdot B_t^{\alpha(1-\xi)} H_t^{-\gamma\alpha}\left[\int_0^1 \left(h_t^i\right)^\theta di\right]^{-\alpha} h_{t+1}^i \\
& + \left[\frac{\alpha A\Psi_c^{1-\xi(1-\alpha)}}{\phi^{\alpha-1}}\right](1-\tau_{t+1})\left(\frac{\tau_t}{1-\tau_t}\right)^{\xi(1-\alpha)} \\
& \cdot B_t^{(\alpha-1)(1-\xi)} H_t^{\gamma(1-\alpha)}\left[\int_0^1 \left(h_t^i\right)^\theta di\right]^{1-\alpha} b_t^i
\end{aligned}
$$

This equation comes from the household's first order condition, where income in period $(t+1)$ is replaced by (8.15) and prices by (8.14) and (8.13).

For $i \in [0,1]$,

$$(8.17) \qquad h_{t+1}^i = \phi H_t^\gamma \left(h_t^i\right)^\theta \left(\frac{\tau_t}{1-\tau_t}\right)^\xi \Psi_c^{-\xi} B_t^\xi$$

$$(8.18) \qquad B_{t+1} = \int_0^1 b_{t+1}^i di$$

$$(8.19) \qquad H_{t+1} = \int_0^1 h_{t+1}^i di$$

$$(8.20) \qquad \frac{(\tau_{t+1})^{\xi\rho-1}}{(1-\tau_{t+1})^{\rho-1}} = \left[\frac{(\beta_c \Psi_b^\rho + \beta_b \Psi_c^\rho)}{\xi\beta_h\phi^\rho}\right] \frac{\left(I_{t+1}^m\right)^\rho}{\left(\int_0^1 I_{t+1}^i di\right)^{\xi\rho}} . H_{t+1}^{-\gamma\rho}\left(h_{t+1}^m\right)^{-\theta\rho}$$

$$(8.21) \qquad r_{t+1} = \left[\frac{\alpha A\Psi_c^{-\xi(1-\alpha)}}{\phi^{\alpha-1}}\right]\left(\frac{\tau_t}{1-\tau_t}\right)^{\xi(1-\alpha)} B_t^{(\alpha-1)(1-\xi)} H_t^{\gamma(1-\alpha)}\left[\int_0^1 \left(h_t^i\right)^\theta di\right]^{1-\alpha}$$

$$(8.22) \qquad w_{t+1} = \left[\frac{(1-\alpha)\, A\Psi_c^{\xi\alpha}}{\phi^\alpha}\right]\left(\frac{\tau_t}{1-\tau_t}\right)^{-\xi\alpha} B_t^{\alpha(1-\xi)} H_t^{-\gamma\alpha}\left[\int_0^1 \left(h_t^i\right)^\theta di\right]^{-\alpha}$$

For $i \in [0,1]$,

$$(8.23) \qquad I_{t+1}^i = w_{t+1} h_{t+1}^i + r_{t+1} b_t^i$$

and agent m is the median voter.

8.6.2. *Steady state solution.* We look solve for a steady state solution using a guess-and-verify strategy.

First, guess that in the steady state:

$$h_{t+1}^{i} = h^{*} > 0 \quad \forall i, t$$

$$b_{t+1}^{i} = b^{*} > 0 \quad \forall i, t$$

Proceed to verify that this proposed solution is consistent with equations (8.16)-(8.23) listed above. From (8.18) and (8.19), in steady state:

$$B_{t+1}^{*} = \int_{0}^{1} b^{*} di = b^{*} \quad \forall t$$

$$H_{t+1}^{*} = \int_{0}^{1} h^{*} di = h^{*} \quad \forall t$$

From (8.12), capital in efficiency units is constant in steady state,

$$\begin{aligned} k_{t+1}^{*} &= \frac{b^{*}}{h^{*}} \\ &= k^{*} \quad \forall t \end{aligned}$$

This implies that prices are constant in steady state from (8.9) and (8.10)

$$\begin{aligned} r_{t+1}^{*} &= \alpha A \left(\frac{b^{*}}{h^{*}} \right)^{\alpha - 1} \\ &= r^{*} \quad \forall t \end{aligned}$$

$$\begin{aligned} w_{t+1}^{*} &= (1 - \alpha) A \left(\frac{b^{*}}{h^{*}} \right)^{\alpha} \\ & \quad w^{*} \quad \forall t \end{aligned}$$

Also, incomes are constant in steady state from (8.23),

$$\begin{aligned} I_{t+1}^{i*} &= w^{*} h^{*} + r^{*} b^{*} \\ &= A \left(b^{*} \right)^{\alpha} \left(h^{*} \right)^{(1-\alpha)} \\ &= I^{*} \quad \forall i, t. \end{aligned}$$

Tax rates are constant since from (8.20),

$$\begin{aligned} \frac{\left(\tau_{t+1} \right)^{\xi \rho - 1}}{\left(1 - \tau_{t+1} \right)^{\rho - 1}} &= \Gamma_{4} \left(b^{*} \right)^{\alpha \rho (1 - \xi)} \left(h^{*} \right)^{\rho [(1-\alpha)(1-\xi) + (\gamma + \theta)]} \\ &= \frac{\left(\tau^{*} \right)^{\xi \rho - 1}}{\left(1 - \tau^{*} \right)^{\rho - 1}} \quad \forall t \end{aligned}$$

where

$$\Gamma_4 = A^{\rho(1-\xi)}\left(\frac{(\beta_c\Psi_b^{\rho}+\beta_b\Psi_c^{\rho})}{\xi\beta_h\phi^{\rho}}\right) > 0.$$

It is left to verify that the proposed steady state solution is consistent with equations (8.16) and (8.17).

From (8.17),

$$\begin{aligned} h^* &= \left[\phi^{\frac{1}{1-\gamma-\theta}}\Psi_c^{-\frac{\xi}{1-\gamma-\theta}}\right]\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi}{1-\gamma-\theta}}(b^*)^{\frac{\xi}{1-\gamma-\theta}} \\ &\equiv \Gamma_3.\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi}{1-\gamma-\theta}}.(b^*)^{\frac{\xi}{1-\gamma-\theta}} \end{aligned}$$

where

$$\Gamma_3 = \left(\phi^{\frac{1}{1-\gamma-\theta}}\Psi_c^{-\frac{\xi}{1-\gamma-\theta}}\right) > 0.$$

From (8.16),

$$0 = b^*\begin{pmatrix} 1-\Gamma_1\left(\frac{1-\tau^*}{\left(\frac{\tau^*}{1-\tau^*}\right)^{\alpha\xi}}\right)(h^*)^{(1-\alpha(\gamma+\theta))}(b^*)^{\alpha(1-\xi)-1} \\ -\Gamma_2\left(\frac{1-\tau^*}{\left(\frac{\tau^*}{1-\tau^*}\right)^{-\xi(1-\alpha)}}\right)(h^*)^{(1-\alpha)(\gamma+\theta)}(b^*)^{-(1-\alpha)(1-\xi)} \end{pmatrix}$$

Putting the above expression for h^* into the above and doing some algebra,

$$0 = b^*\begin{pmatrix} 1-\Gamma_1\Gamma_3^{(1-\alpha\gamma-\alpha\theta)}(1-\tau^*)\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi(1-\alpha)}{1-\gamma-\theta}}(b^*)^{\frac{\xi(1-\alpha)}{1-\gamma-\theta}-(1-\alpha)} \\ -\Gamma_2.\Gamma_3^{(1-\alpha)(\gamma+\theta)}(1-\tau^*)\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi(1-\alpha)}{1-\gamma-\theta}}(b^*)^{\frac{\xi(1-\alpha)}{1-\gamma-\theta}-(1-\alpha)} \end{pmatrix}$$

where,

$$\Gamma_1 = \frac{(1-\alpha)\,A\Psi_c^{1+\alpha\xi}}{\phi^{\alpha}} > 0$$

$$\Gamma_2 = \frac{\alpha A\Psi_c^{1-\xi(1-\alpha)}}{\phi^{\alpha-1}} > 0.$$

If a positive solution exists to the preceding equation for b^* then we are done. The $b^* = 0$ solution is ruled out. Therefore, we need a solution for

(8.24)

$$0 = 1-\left[\Gamma_1\Gamma_3^{(1-\alpha\gamma-\alpha\theta)}+\Gamma_2.\Gamma_3^{(1-\alpha)(\gamma+\theta)}\right](1-\tau^*).\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi(1-\alpha)}{1-\gamma-\theta}}(b^*)^{\frac{\xi(1-\alpha)}{1-\gamma-\theta}-(1-\alpha)}$$

The above system is extremely difficult to solve.[13] The main reason is that, with majority voting to determine the tax rate, the dynamic system is highly non-linear. This is problematic because the income distribution is not stationary. Indeed, it is straightforward to show that, even with initial differences in bequest levels, the

[13]Notes with further analysis that clarify this difficulty are available from the authors.

long-run properties of the system imply that bequest levels converge to the same level across groups. Differences in spillovers, formalized by differences in H_{jt}, or other initial level differences will not lead to long run differences in human capital levels across groups.

8.6.3. *Understanding long run differences in preferred tax rates.* The empirical results suggest that differences in circumstances across racial, gender, and class groups continue to render themselves salient in individuals' determination of their preferred levels of income redistribution. One possible explanation for this is differences in the returns to educational expenditure across these identity groups.

To demonstrate this formally, it is simplest to assume that actual tax rates are determined exogenously to the model, and to allow for a degenerate distribution of initial endowments within each identity group at a single level. Each agent will still have a preferred tax rate, even if she does not participate in a voting mechanism. Each agent belongs to an identity group j; and each identity group has a specific return to educational expenditure, ξ_j. The goal is to determine how a preferred tax rate varies with ξ_j in the long-run of the model.

To determine how the preferred tax rate varies with ξ_j in the long-run of the model, start with the expression for b_j^*, which will apply to each identity group j, using (8.24):

(8.25)

$$0 = 1 - \left[\Gamma_{1j}\Gamma_{3j}^{(1-\alpha\gamma-\alpha\theta)} + \Gamma_{2j}.\Gamma_{3j}^{(1-\alpha)(\gamma+\theta)}\right](1-\tau^*).\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi_j(1-\alpha)}{1-\gamma-\theta}}\left(b_j^*\right)^{\frac{\xi_j(1-\alpha)}{1-\gamma-\theta}-(1-\alpha)}$$

Rewriting to solve for b_j^*,

$$b_j^* = \left[\left[\Gamma_{1j}\Gamma_{3j}^{(1-\alpha\gamma-\alpha\theta)} + \Gamma_{2j}.\Gamma_{3j}^{(1-\alpha)(\gamma+\theta)}\right](1-\tau^*)\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi_j(1-\alpha)}{1-\gamma-\theta}}\right]^{\frac{-(1-\gamma-\theta)}{(1-\alpha)(\xi_j+\gamma+\theta-1)}}.$$

How does b_j^* vary with the exogenously determined tax rate, τ^*? To answer this question, consider one term on the right hand side at a time in the above expression.

Rewrite the first term as:

$$\left[\Gamma_{1j}\Gamma_{3j}^{(1-\alpha\gamma-\alpha\theta)} + \Gamma_{2j}.\Gamma_{3j}^{(1-\alpha)(\gamma+\theta)}\right]^{\frac{-(1-\gamma-\theta)}{(1-\alpha)(\xi_j+\gamma+\theta-1)}} = \left[A\phi^{\frac{1-\alpha}{1-\gamma-\theta}}\right]^{\frac{-(1-\gamma-\theta)}{(1-\alpha)(\xi_j+\gamma+\theta-1)}}\left[\Psi_c\right]^{\frac{\xi_j-\frac{(1-\gamma-\theta)}{(1-\alpha)}}{\xi_j-(1-\gamma-\theta)}}.$$

Take the derivative of this term with respect to ξ_j to get

$$\left[A\phi^{\frac{(1-\alpha)}{1-\gamma-\theta}}\right]^{\frac{-(1-\gamma-\theta)}{(1-\alpha)\left(\xi_j+\gamma+\theta-1\right)}}\left[\Psi_c\right]^{\frac{\xi_j-\frac{(1-\gamma-\theta)}{(1-\alpha)}}{\xi_j-(1-\gamma-\theta)}}\left\{\begin{array}{c}\ln\left[A\phi^{\frac{1-\alpha(\gamma+\theta)}{1-\gamma-\theta}}\right]\left[\frac{(1-\gamma-\theta)}{(1-\alpha)\left(\xi_j+\gamma+\theta-1\right)^2}\right]\\ +\left[\ln\Psi_c\right]\frac{(1-\gamma-\theta)\left[\frac{1}{(1-\alpha)}-1\right]}{\left[\xi_j-(1-\gamma-\theta)\right]^2}\end{array}\right\}.$$

This expression is negative if $\left[A\phi^{\frac{(1-\alpha)}{1-\gamma-\theta}}\right]<1$. The productivity measure A may be expressed in units such that this is the case.

Next, rewrite the second term:

$$\left(1-\tau^*\right)^{\frac{-(1-\gamma-\theta)}{(1-\alpha)\left(\xi_j+\gamma+\theta-1\right)}}.$$

Taking the derivative of this term with respect to ξ_j, we get

$$\left(1-\tau^*\right)^{\frac{-(1-\gamma-\theta)}{(1-\alpha)\left(\xi_j+\gamma+\theta-1\right)}}\left[\ln\left(1-\tau^*\right)\right]\frac{(1-\gamma-\theta)}{(1-\alpha)\left(\xi+\gamma+\theta-1\right)^2}<0.$$

Lastly, rewrite the third term:

$$\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{-\xi_j}{\left(\xi_j+\gamma+\theta-1\right)}}.$$

Taking the derivative of this term with respect to ξ_j, we get

$$\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{-\xi_j}{\left(\xi_j+\gamma+\theta-1\right)}}\left[\ln\left(\frac{\tau^*}{1-\tau^*}\right)\right]\frac{(1-\gamma-\theta)}{\xi^2}\left[\frac{1}{1+\frac{(1-\gamma-\theta)}{\xi}}\right]^2$$

This expression is negative if the fixed τ^* is less than 0.5.

Thus, under reasonably general conditions, $\frac{\partial b_j^*}{\partial\xi_j}<0$.

Next, recall the expression for h_j^* :

$$h_j^*=\Gamma_{3j}\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi_j}{1-\gamma-\theta}}\left(b_j^*\right)^{\frac{\xi_j}{1-\gamma-\theta}}$$

Again, the question is how does h_j^* vary with the exogenously determined tax rate, τ^*? To answer this question, consider one term on the right hand side at a time in the above expression.

First, rewrite the second term

$$\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi_j}{1-\gamma-\theta}}$$

Taking the derivative of this term with respect to ξ_j, we get

$$\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi_j}{1-\gamma-\theta}}\ln\left(\frac{\tau^*}{1-\tau^*}\right)\frac{1}{1-\gamma-\theta}$$

This expression is negative if the fixed τ^* is less than 0.5.

Next, combine the first and third terms:

$$\begin{aligned}
& \Gamma_{3j}\left(b_j^*\right)^{\frac{\xi_j}{1-\gamma-\theta}} \\
= & \left[\phi^{\frac{1}{1-\gamma-\theta}}\left(\Psi_c\right)^{\frac{-\xi_j}{1-\gamma-\theta}}\right]\left[A\phi^{\frac{(1-\alpha)}{1-\gamma-\theta}}\right]^{\frac{-\xi_j}{(1-\alpha)(\xi+\gamma+\theta-1)}} \cdot \\
& \left[\Psi_c\right]^{\frac{\xi_j-\frac{(1-\gamma-\theta)}{(1-\alpha)}}{\xi_j-(1-\gamma-\theta)}\frac{\xi_j}{1-\gamma-\theta}}\left(1-\tau^*\right)^{\frac{-\xi_j}{(1-\alpha)(\xi_j+\gamma+\theta-1)}}\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi_j}{(\xi_j+\gamma+\theta-1)}\frac{\xi_j}{1-\gamma-\theta}} \\
= & \phi^{\frac{1}{1-\gamma-\theta}}\left[A\phi^{\frac{(1-\alpha)}{1-\gamma-\theta}}\right]^{\frac{-\xi_j}{(1-\alpha)(\xi_j+\gamma+\theta-1)}}\left[\Psi_c\right]^{\frac{(1-\gamma-\theta)\left[1-\frac{1}{(1-\alpha)}\right]}{\xi_j-(1-\gamma-\theta)}\frac{\xi_j}{1-\gamma-\theta}} \\
& \cdot\left(1-\tau^*\right)^{\frac{-\xi_j}{(1-\alpha)(\xi_j+\gamma+\theta-1)}}\left(\frac{\tau^*}{1-\tau^*}\right)^{\frac{\xi_j^2}{(\xi_j+\gamma+\theta-1)(1-\gamma-\theta)}}
\end{aligned}$$

The first term involving ξ_j is decreasing in ξ_j if $\left[A\phi^{\frac{(1-\alpha)}{1-\gamma-\theta}}\right] < 1$ and $2\xi_j + \gamma + \theta - 1 > 0$. The second term and third terms are decreasing in ξ_j if $\gamma + \theta - 1 < 0$. The fourth term is decreasing in ξ_j if $\tau^* < 0.5$ and $\xi_j + 2(\gamma + \theta - 1) < 0$.

Thus, under these conditions, $\frac{\partial h_j^*}{\partial \xi_j} < 0$.

Now we examine each group's preferred tax rate, noting that the actual tax rate is exogenously determined.

$$\frac{(\hat{\tau}_j)^{\xi_j\rho-1}}{(1-\hat{\tau}_j)^{\rho-1}} = \left[\frac{(\beta_c\Psi_b^\rho + \beta_b\Psi_c^\rho)}{\xi_j\beta_h\phi^\rho}\right]\frac{\left(I_j^*\right)^\rho}{\left(\sum_j I_j^*\right)^{\xi_j\rho}}\left(H^*\right)^{-\gamma\rho}\left(h_j^*\right)^{-\theta\rho}$$

On the left hand side, note that if ξ_j increases, the curve in $\hat{\tau}_j$ will shift downward (since $\hat{\tau}_j < 1$). On the right hand side, note that if ξ_j increases, $\frac{1}{\xi_j}$ decreases, and $\left(\sum_j I_j^*\right)^{-\xi_j\rho}$ decreases (since $\sum_j I_j^* > 1$ is assumed). On the right hand side, note that if ξ_j increases, I_j^* decreases by the analysis above. On the right hand side, note that if ξ_j increases, $\left(h_j^*\right)^{-\theta\rho}$ increases by the analysis above.

In sum, the effect on the right hand side is ambiguous, though it is overall probably decreasing in ξ_j. See Figure 27. The net effect on the preferred tax rate is ambiguous if the right hand side is decreasing in ξ_j. Probably as ξ_j increases, the preferred tax rate decreases (see point 0 to 3 in the figure). To see why, note that the effect on the left hand side is toward decreasing the preferred tax rate (see point 0 to 1 in the figure). The effect on the right hand side is toward increasing

the preferred tax rate if the right hand size is decreasing in ξ_j (see point 0 to 2 in the figure).

Interpretation of our empirical results is driven by persistent differences in ξ_j across identity groups. A related theoretical model by Breen and Garcia-Peñalosa [10] shows that such persistent differences are possible in an environment where agents use Bayesian updating to form views on returns to human capital investment. In their framework, agents only use information from agents in their identity group, which is assumed to be defined by gender. A simpler possibility is that agents observe outcomes of previous generations in their identity group, back out what optimizing behavior must have been of the previous generation, and then behave the same way. In other words, it is assumed that all information of the current generation agents comes from the outcomes of the previous generation agents in the same identity group.

This assumption is an abstraction, but allows us to make our point simply. The idea that agents use identity, and observe outcomes of previous generations to form a view on what is optimizing behavior, is used by others. For instance, Manski [24] examines the consequences for estimation of allowing returns to education vary across identity groups. He reviews the small econometric literature that allows expected educational returns to vary across identity groups.[14]

Loury [23] suggests that a focus on the supply side is needed in order to understand the skills gap between racial groups. His argument is easily extended to the other identity groups that are found to be salient in this paper's empirical results. Loury argues that people are 'socially located' - they are part of social and cultural networks that have strong influence on behavior via social and psychological forces. These social effects can include role model effects, and the psychological effects can include a self-fulfilling pessimism or optimism.

In terms of the model above, h_t^i is determined by H_{t-1}, e_{t-1} and h_{t-1}^i, which are taken as given by both the t and $t-1$ generations of dynasty i. Thus, in some sense, the human capital level of each generation is not chosen by that generation. However, suppose that the parameter ξ_j varies according to effort combined with

[14]The identity groups are defined by the econometricians. Manski shows that there are serious estimation consequences for defining those groups differently than the individuals. That consideration, while important, is not the concern of this study.

educational expenditure. Suppose that the model for determining ξ_j is that each individual chooses effort by observing the outcomes of the previous generation. Thus, the return to educational expenditure of an individual in identity group j in generation t, ignoring variation within identity groups, will be given by:

$$\xi_{jt} = \frac{\ln h_{jt} - \ln \phi - \gamma \ln H_t - \theta \ln h_{j(t-1)}}{\ln e_{t-1}}. \tag{8.26}$$

Suppose there are exogenous differences in opportunity to earn returns to education that lead to initial differences in ξ_{j0} across j. These differences could be the result of the legacy of slavery, earlier demand-side discrimination against different racial, gender, and class groups, or social norms regarding educational opportunity for the different identity groups. Then, the expression (8.26) above reduces such that $\xi_{jt} = \xi_{j0}\ \forall j, t$.

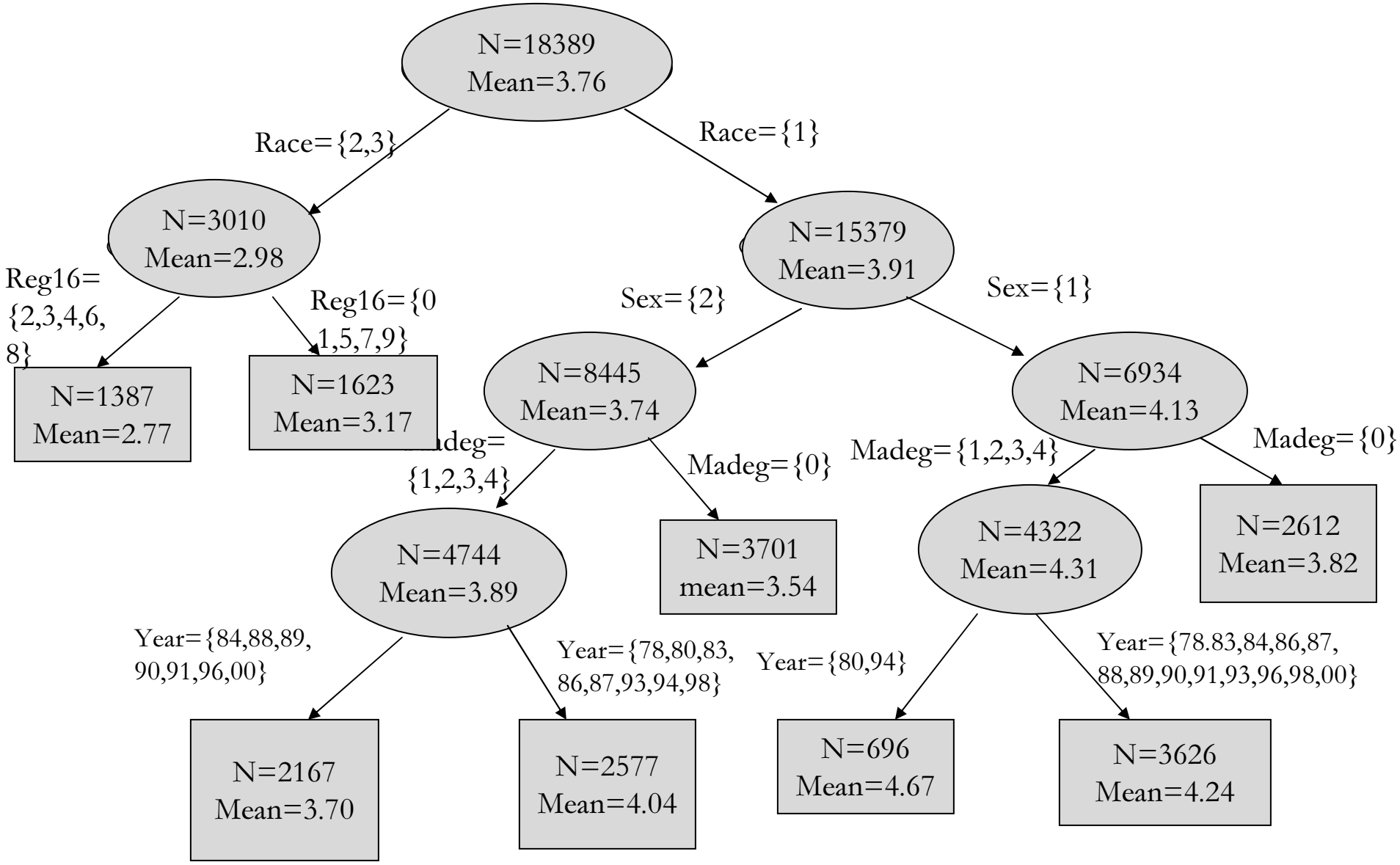
Nothing is said about whether use of this rule to determine effort affecting return to educational expenditure is correct. Indeed, identity can be interpreted here as creating a market failure, as Loury [23] suggests. Strong social and psychological segregation of identity groups can lead to a selective use of information from the previous generation in determining one's own return to education, or other human capital investment. Social location of agents may lead to the use of identity markers to extract information concerning prospects and optimal behavior.

University of Wisconsin and Brookings Institution

E-mail address: lkeely@ssc.wisc.edu

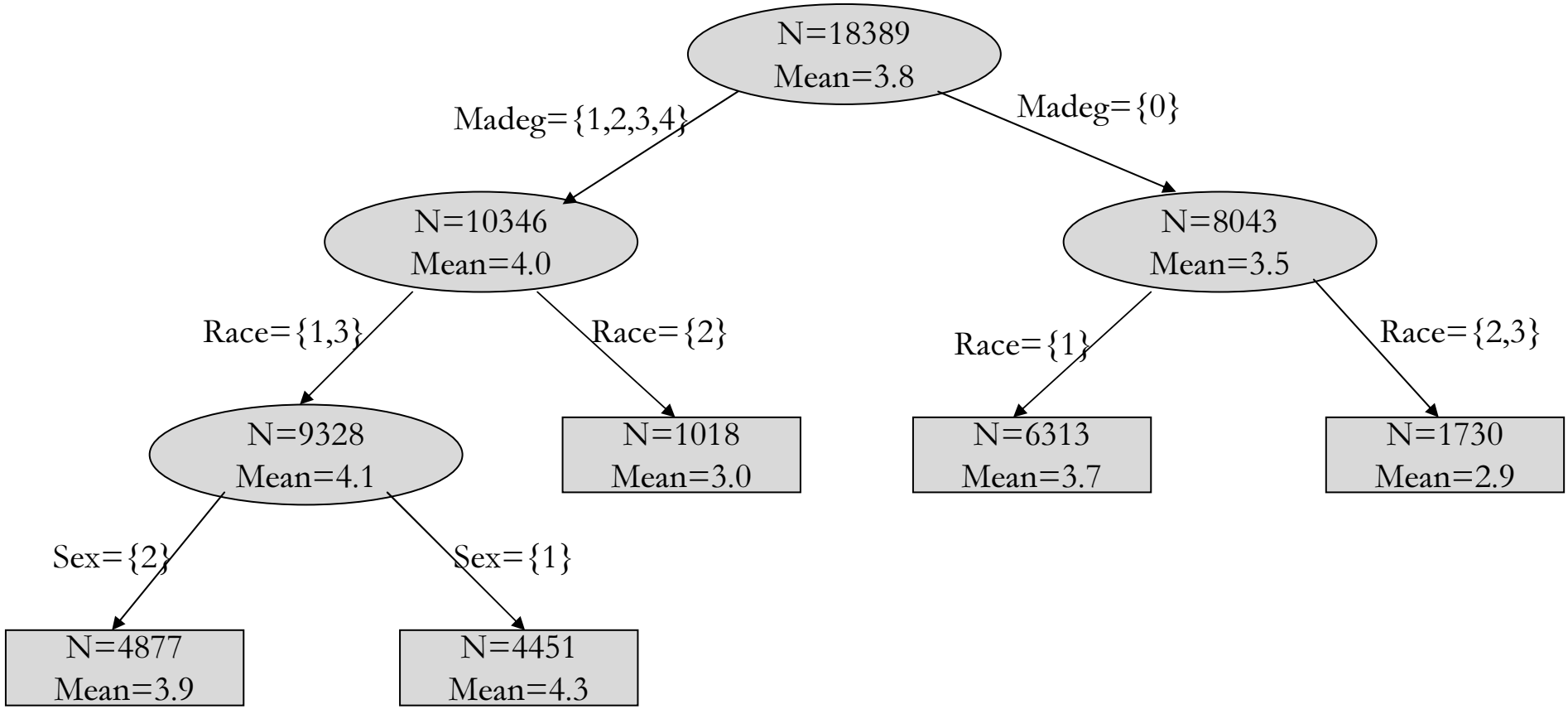
University of Wisconsin

E-mail address: chihmingtan@wisc.edu



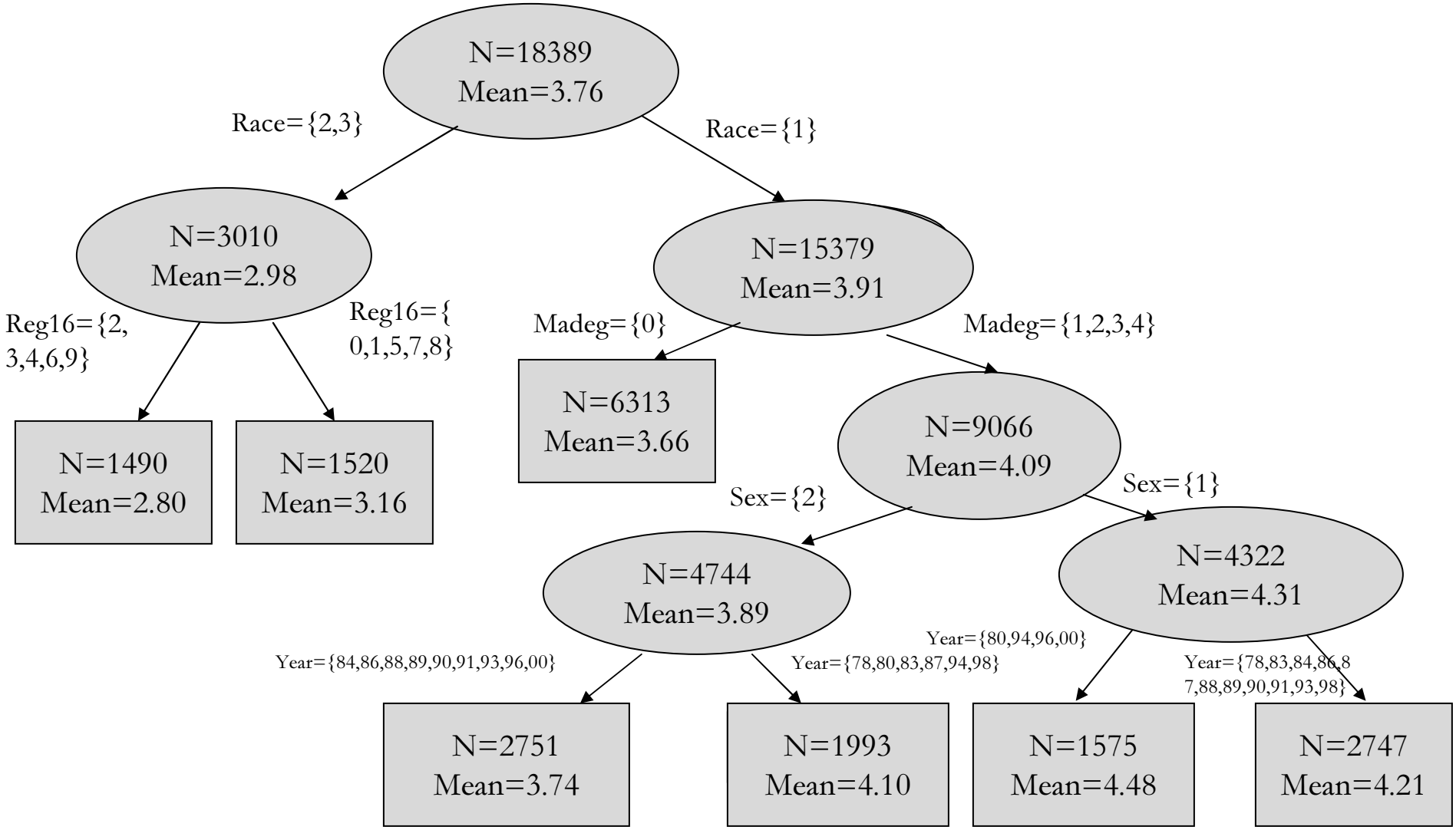
1978-2000 Bayesian Treed Regression

Figure 1



1978-2000 Guide Treed Regression

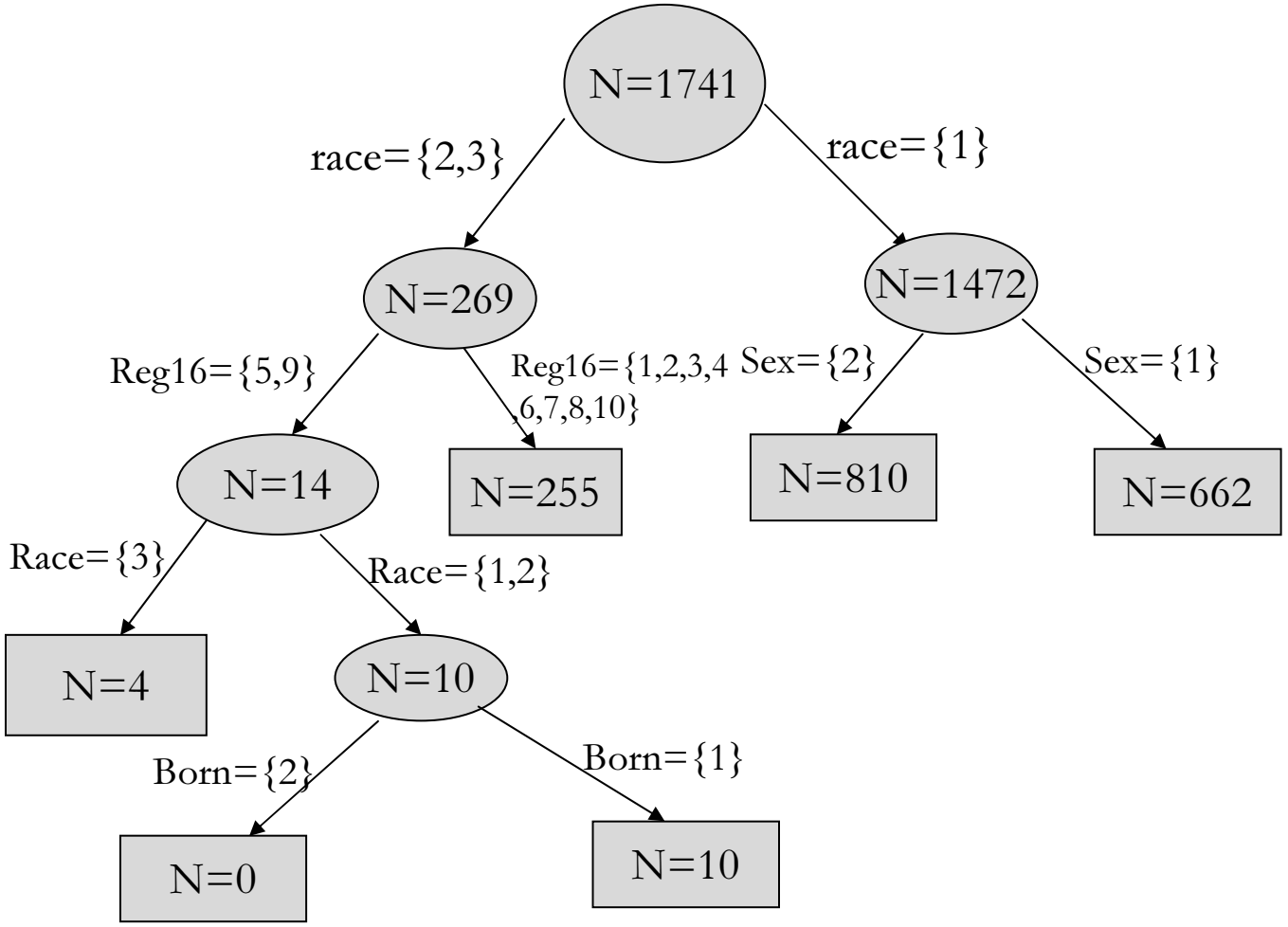
Figure 2



1978-2000 Bayesian Treed Logit Regression A*

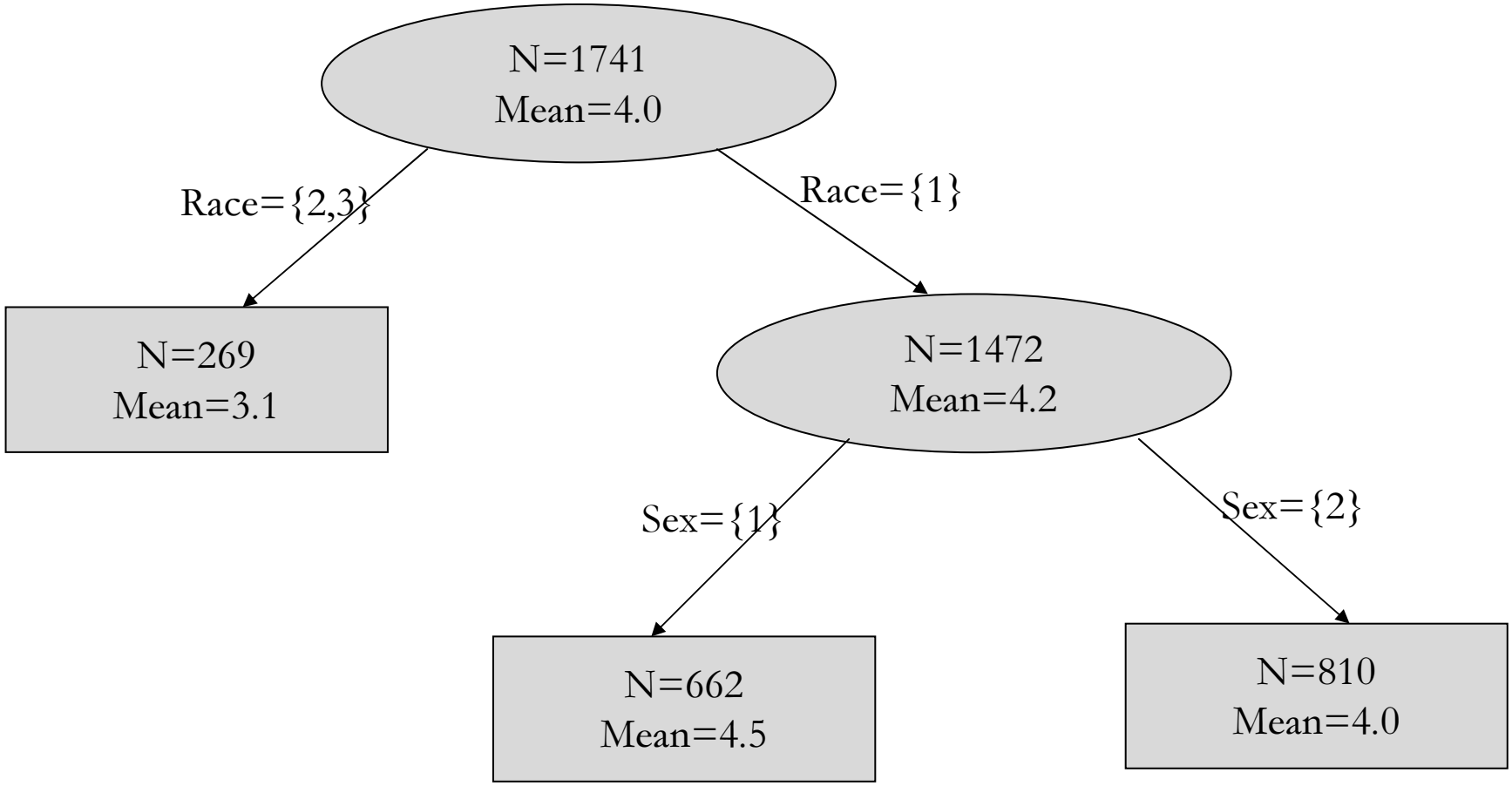
* EQWLTH recoded such that 1-3 is 0 and 4-7 is 1

Figure 3



1994 Bayesian Treed Regression

Figure 4



1994 Guide Treed Regression

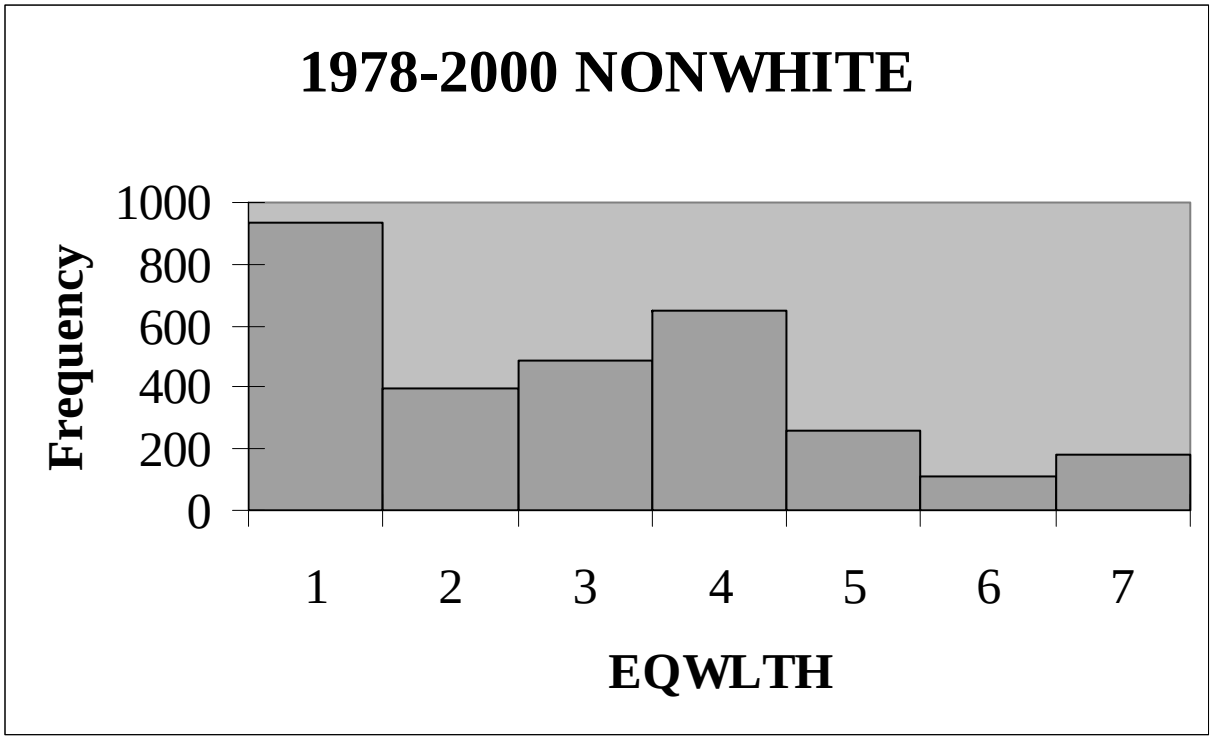
Figure 5

Bayesian Treed Likelihood Ratio tests

Variable	Tree Specifications	Log (L(best tree)/N)	Log (L(tree size=1)/N)	LR	best tree size	critical value	test result
EQWLTH	78-00	5618.3	5138.6	959.4	8	2.17	cannot reject alternative
EQWLTH	94	553.9	508	91.8	5	0.71	cannot reject alternative
FINRELA	78-00	21435.7	20665.2	1541	15	6.57	cannot reject alternative
FINRELA	94	1887.7	1847.3	80.8	7	1.65	cannot reject alternative

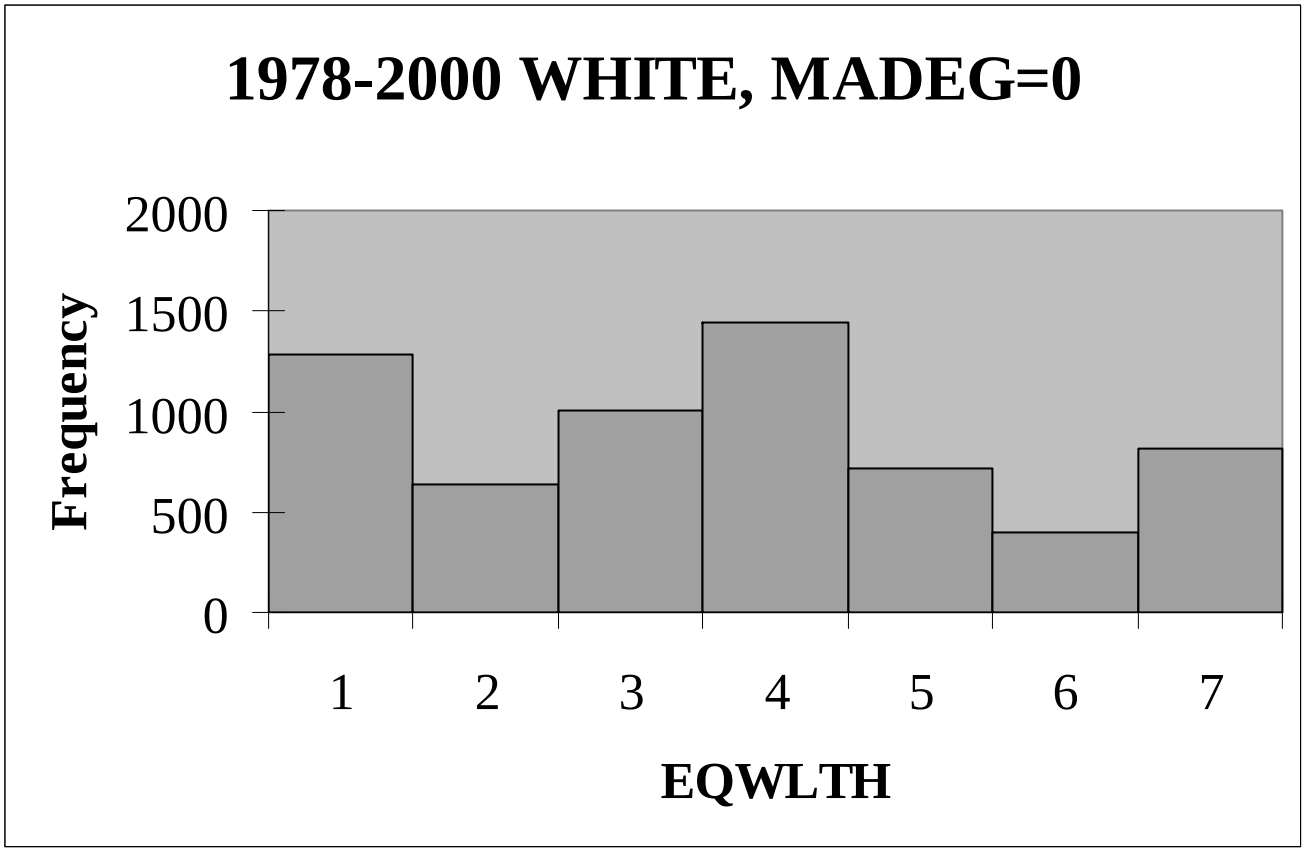
Notes:
N is a normalization which cancels out in calculation of LR statistic
LR statistic is chi-squared distributed with degrees of freedom equal to (number of restrictions is number of nodes to get to best tree from tree size 1) minus 1
Null hypothesis: tree size one is correct model; Alternative hypothesis: best tree size implied model is correct
LR=2log (L(best tree)/L(tree size=1))
citiical value is calculated for 95% significance

Figure 6



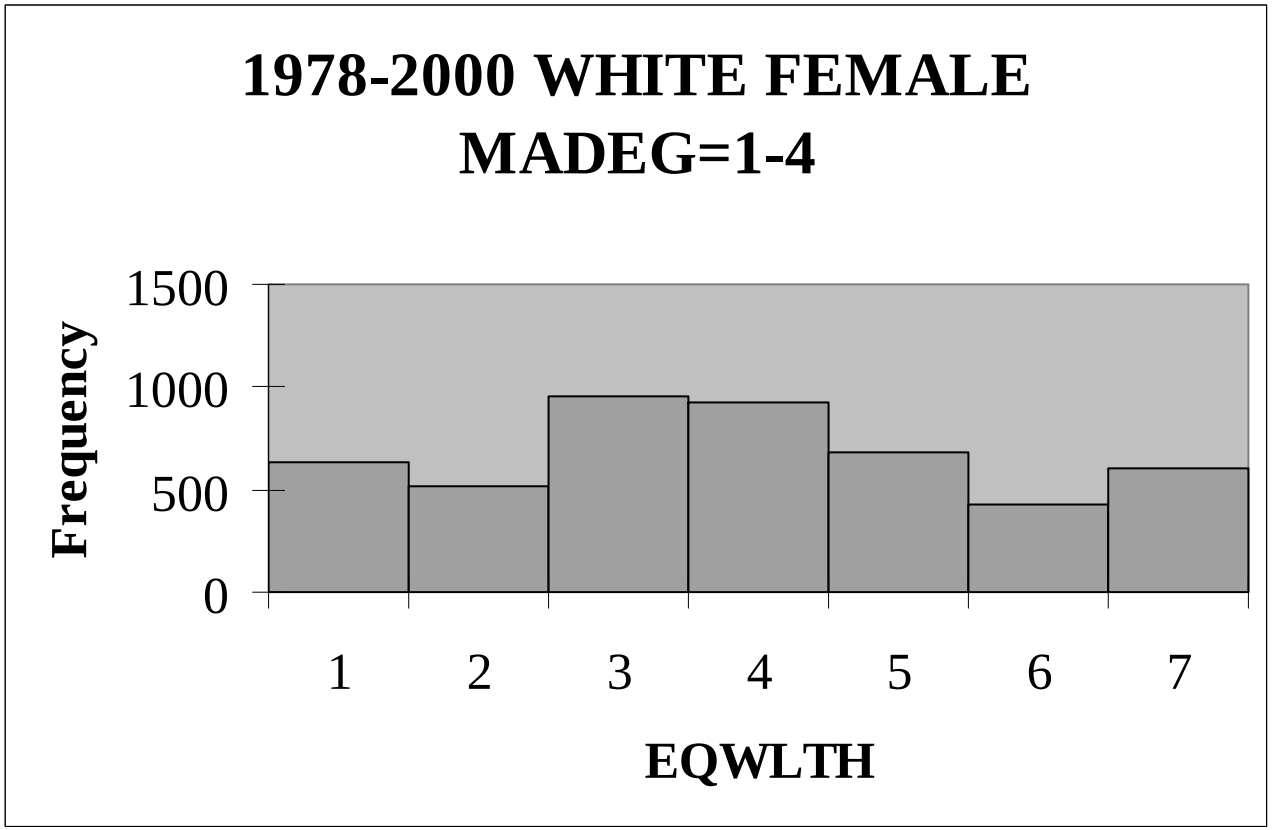
1978-2000 EQWLTH Race=2,3

Figure 7



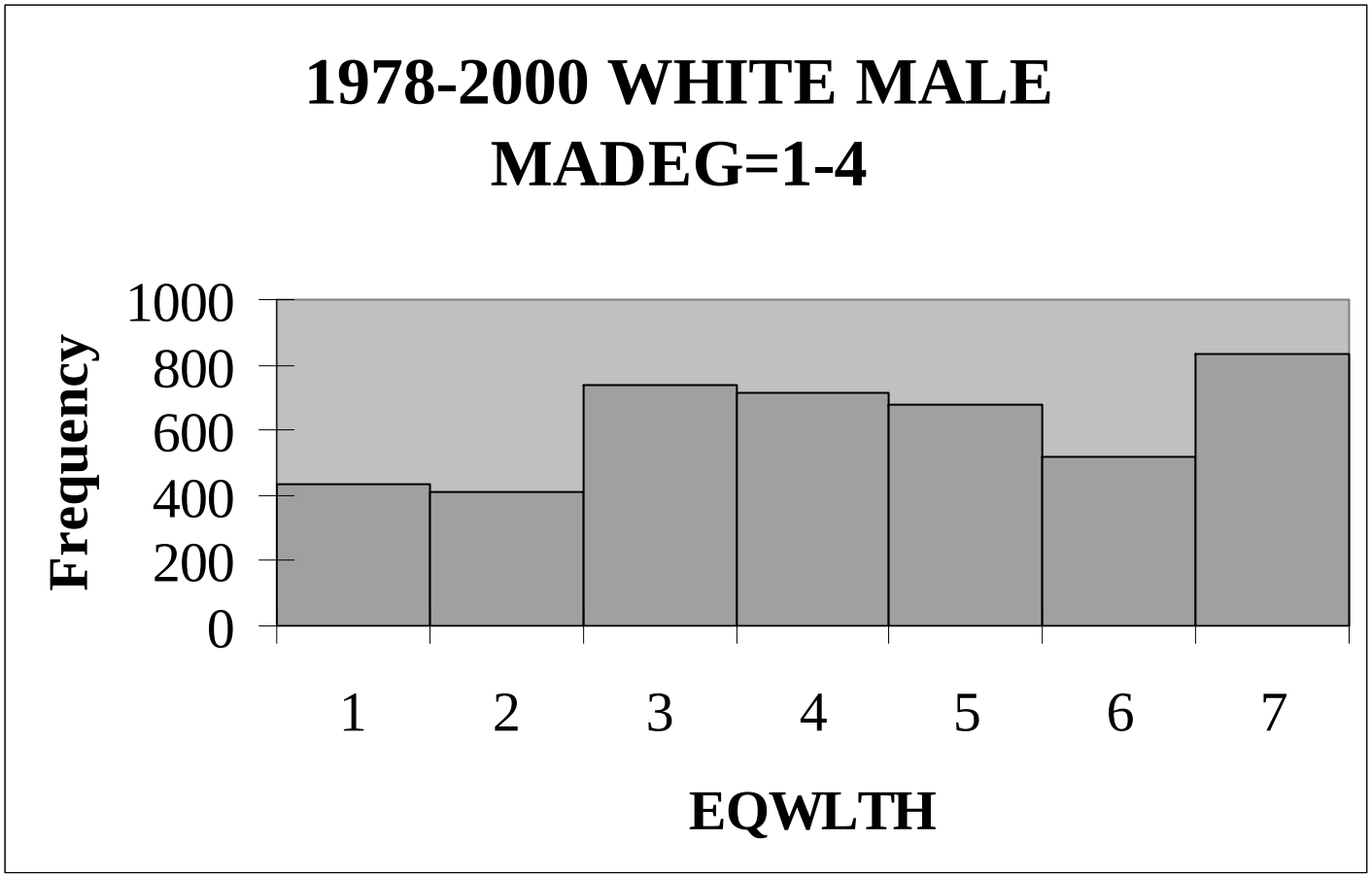
1978-2000 EQWLTH Race=1; Madeg=0

Figure 8



1978-2000 EQWLTH Race=1; Madeg=1-4, Sex=2

Figure 9



1978-2000 EQWLTH Race=1; Madeg=1-4, Sex=1

Figure 10

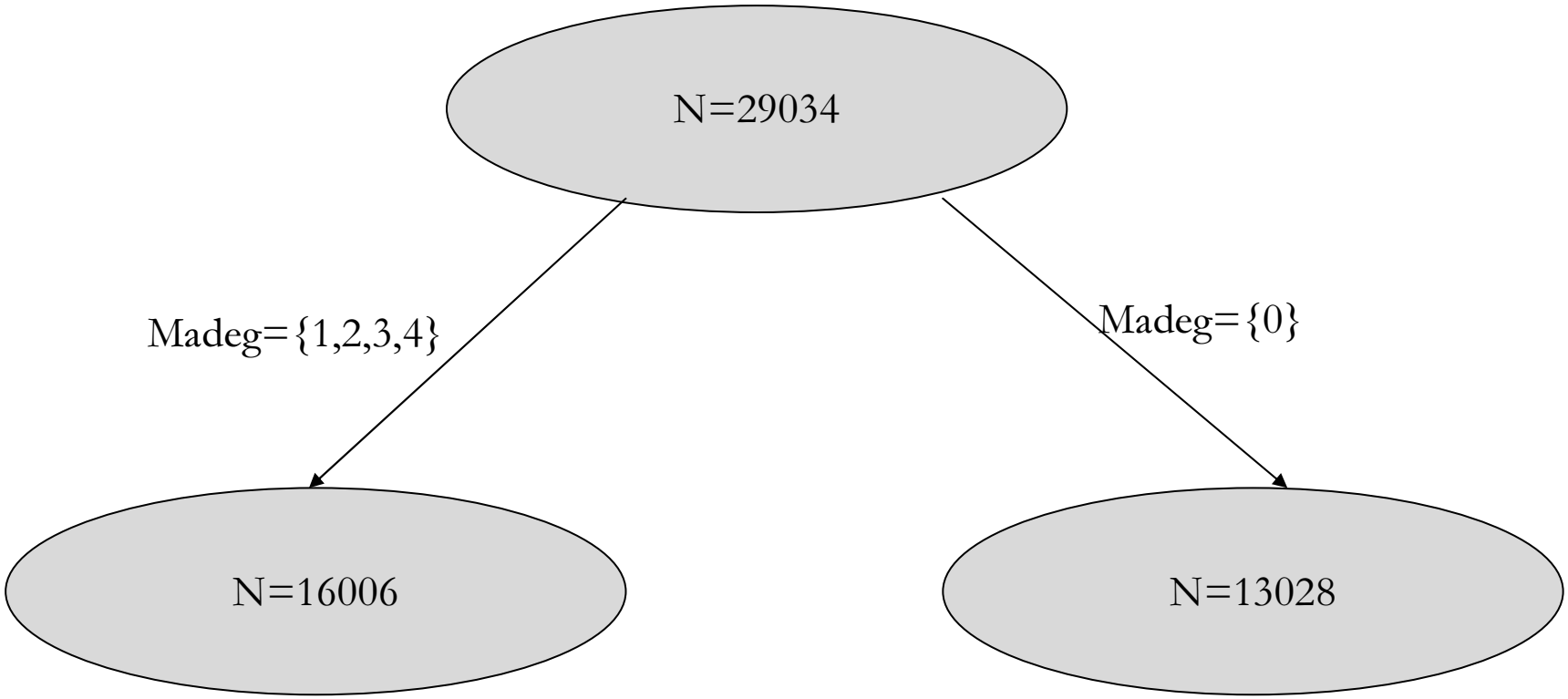


Figure 11

1977-2000 Bayesian Treed Regression-1

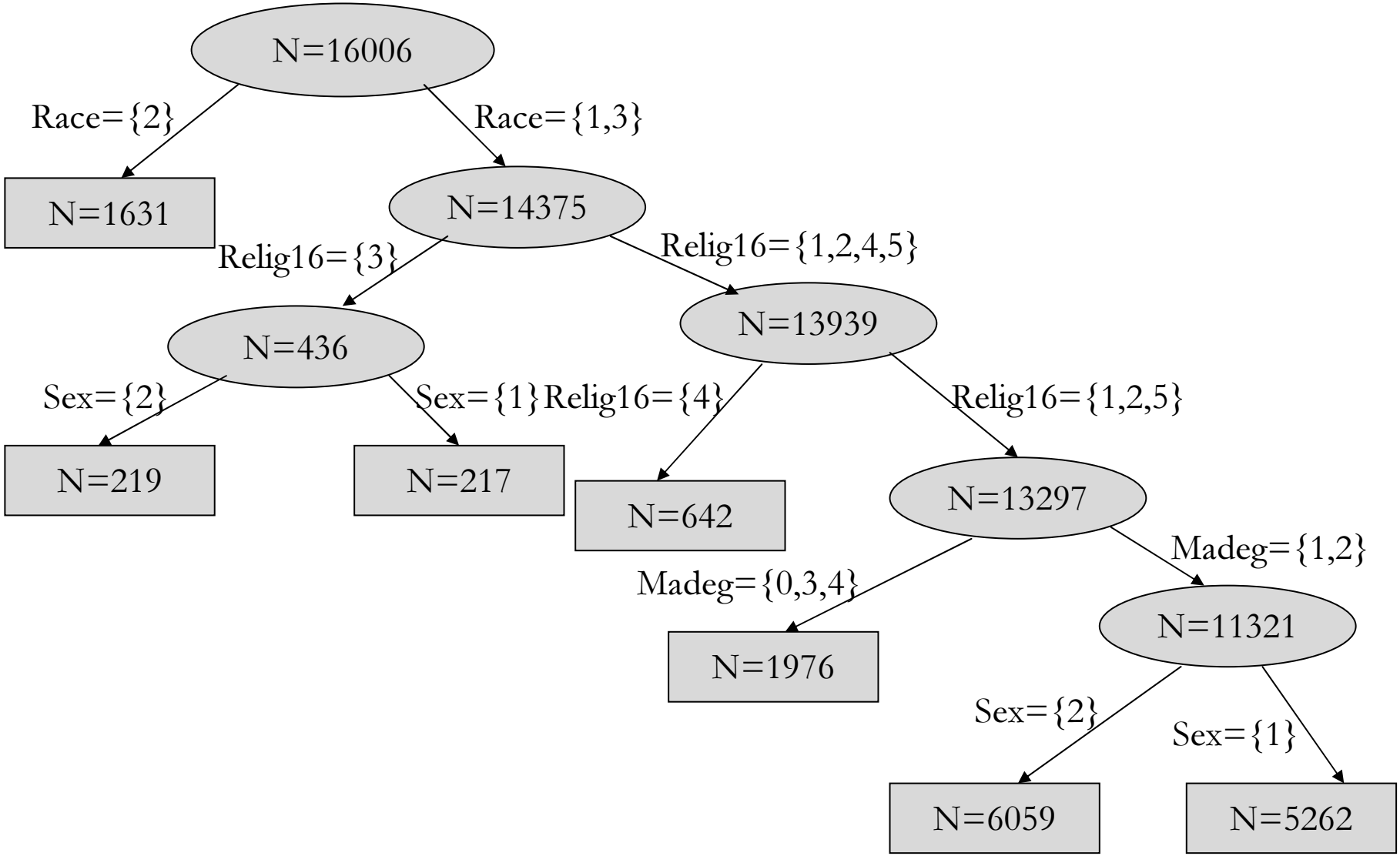

N=16006
Race={2}
Race={1,3}
N=1631
N=14375
Relig16={3}
Relig16={1,2,4,5}
N=436
N=13939
Sex={2}
Sex={1}
Relig16={4}
Relig16={1,2,5}
N=219
N=217
N=642
N=13297
Madeg={1,2}
Madeg={0,3,4}
N=11321
N=1976
Sex={2}
Sex={1}
N=6059
N=5262

Figure 12

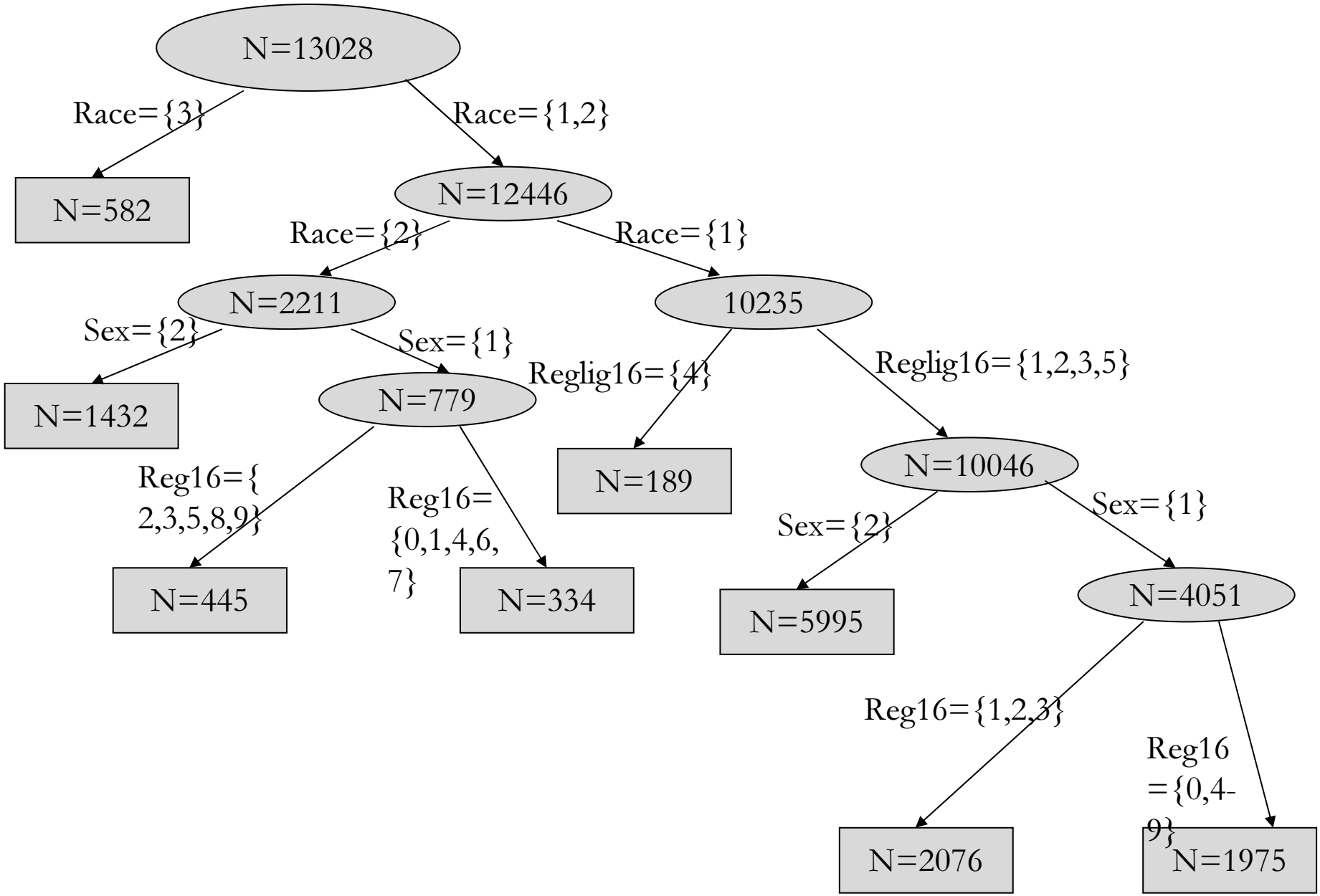


Figure 13

1977-2000 Bayesian Treed Regression-3

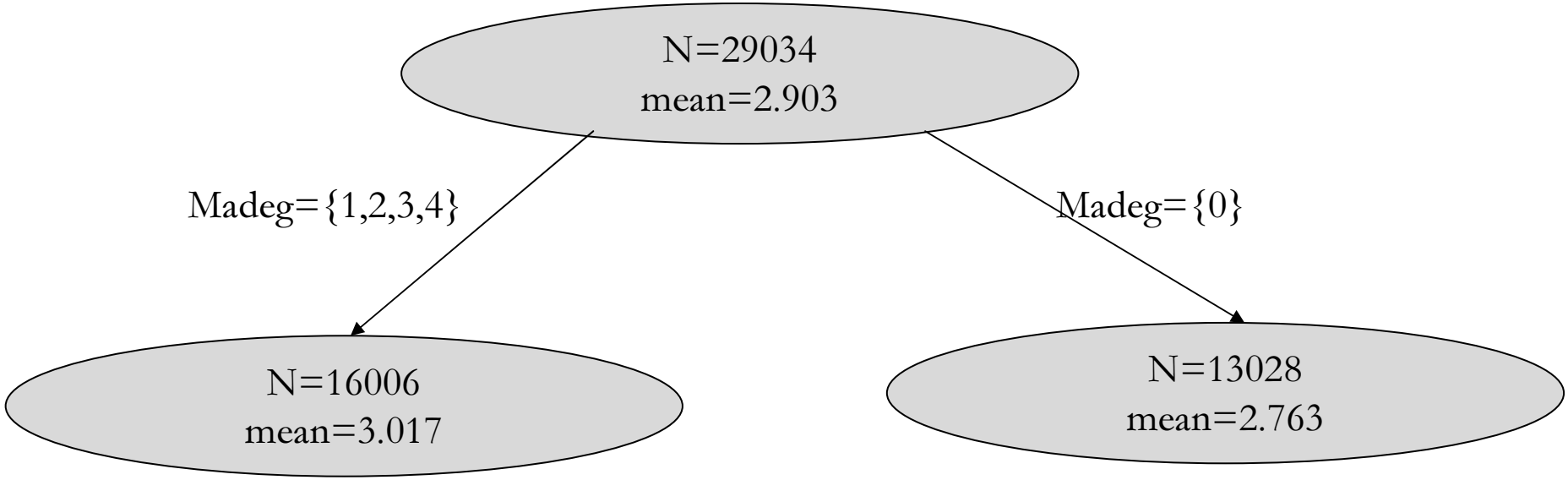


Figure 14

1977-2000 Guide Treed Regression-1

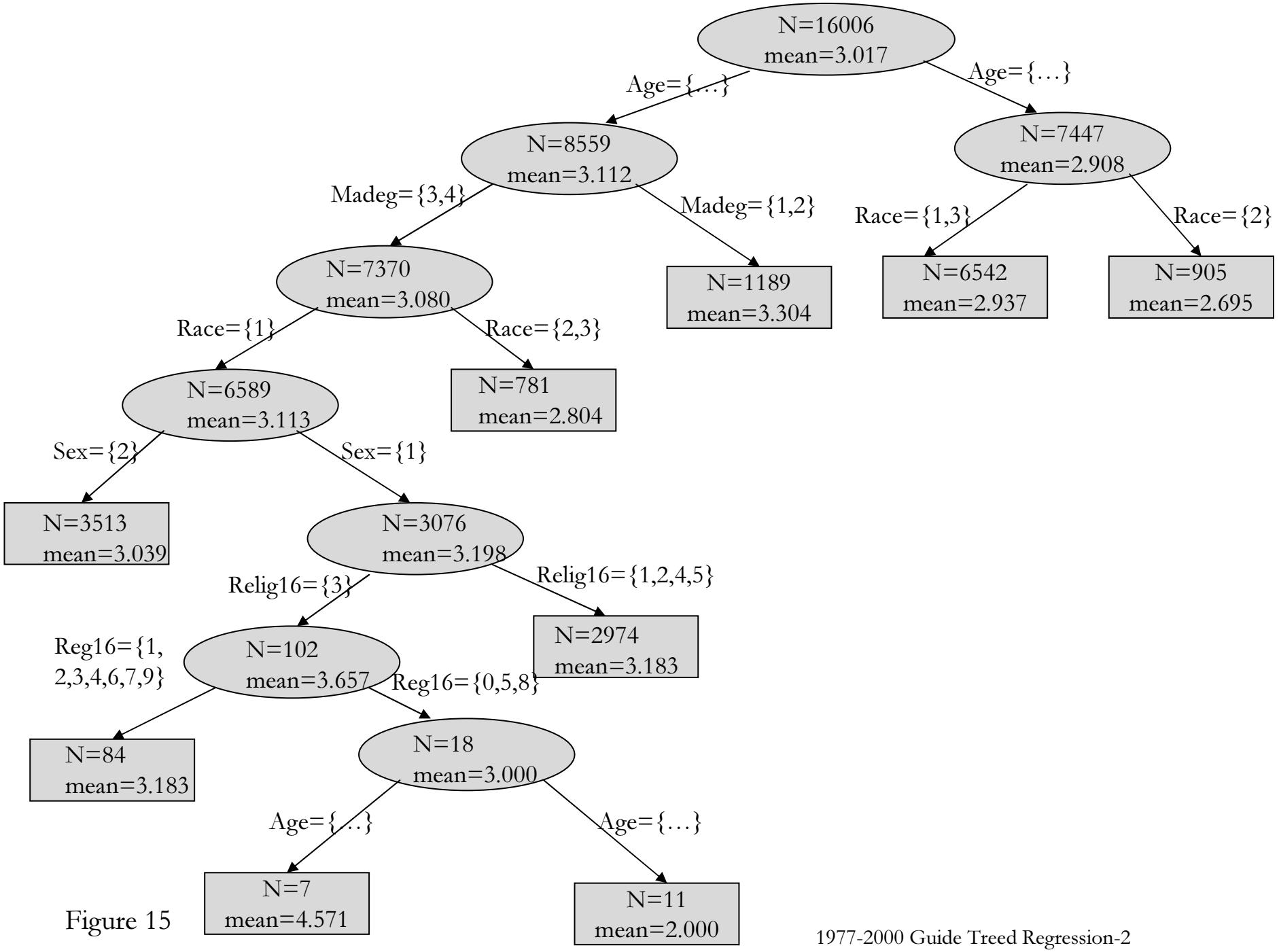


Figure 15

1977-2000 Guide Treed Regression-2

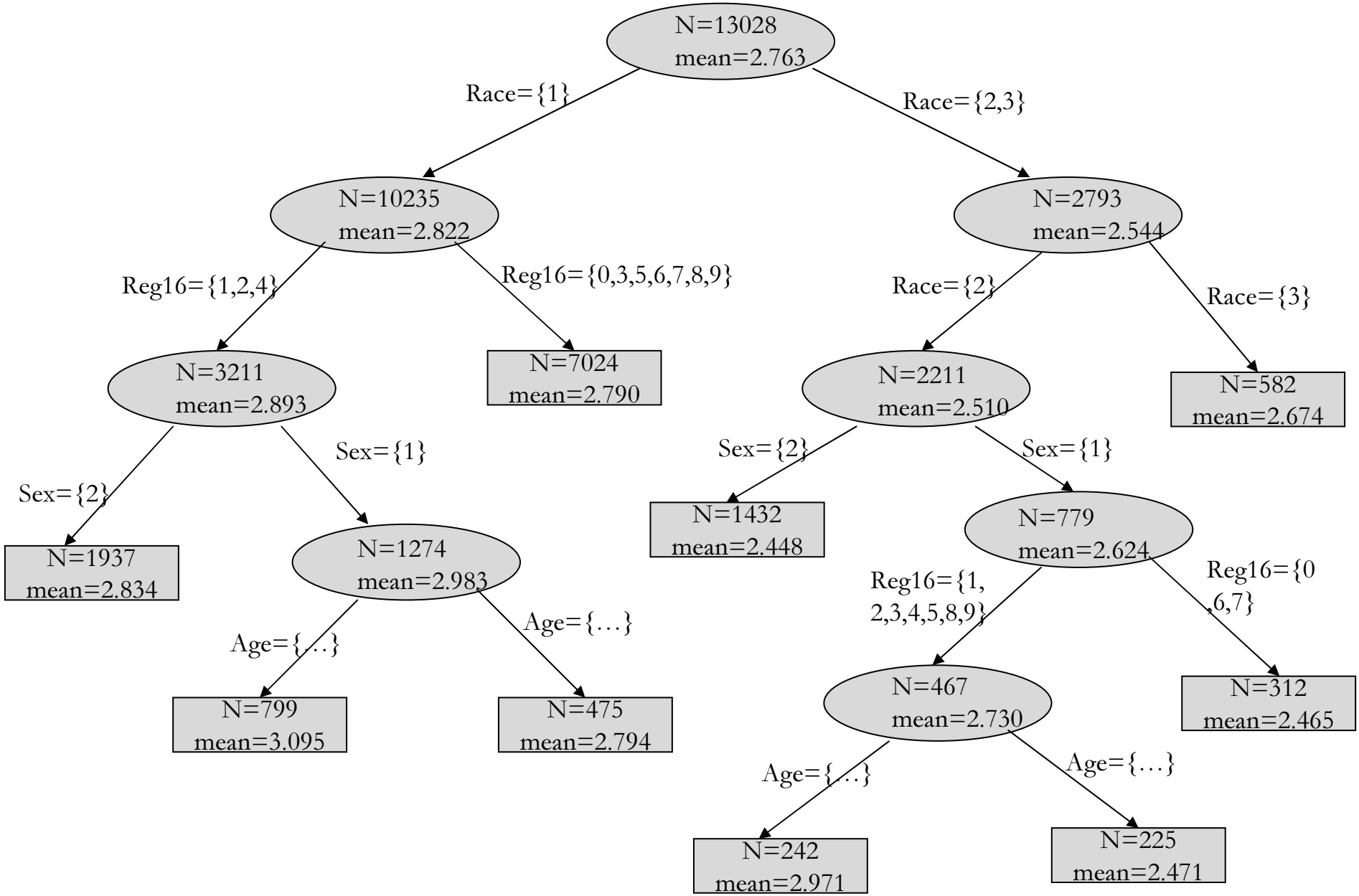
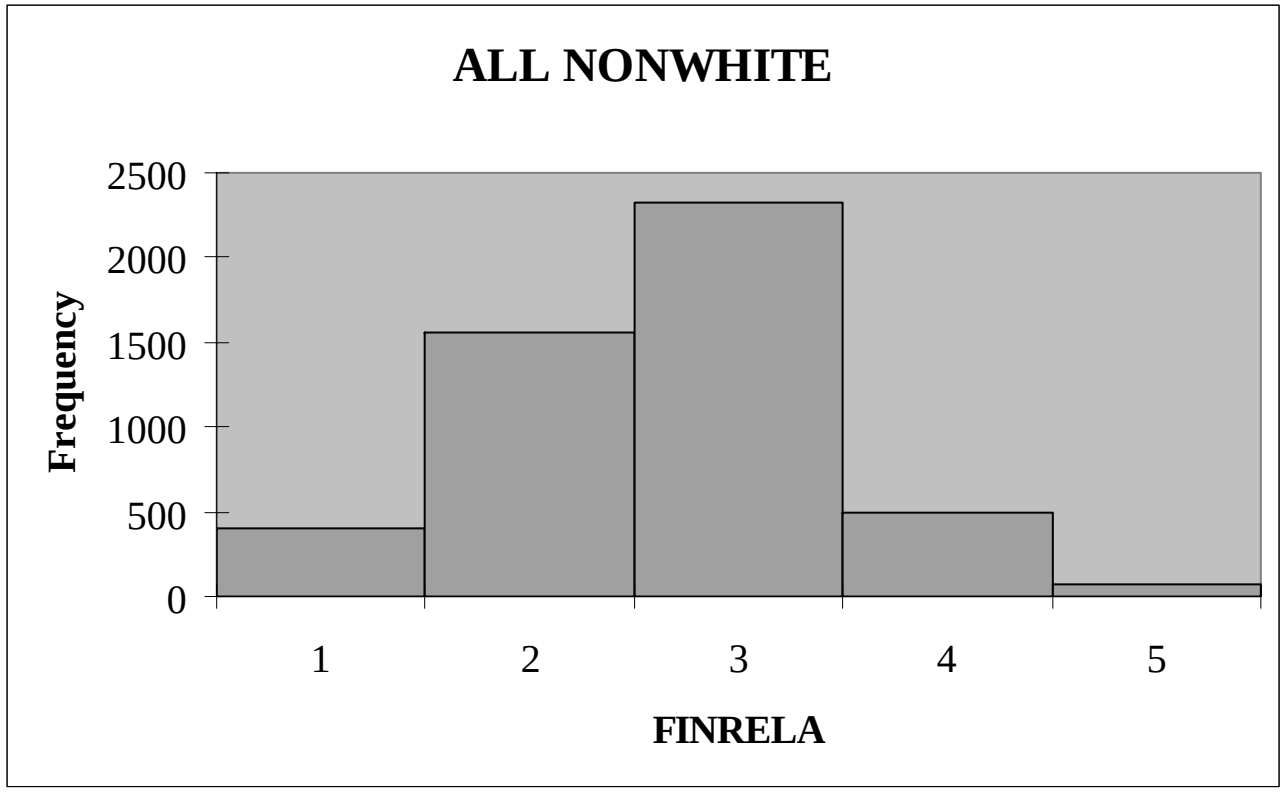


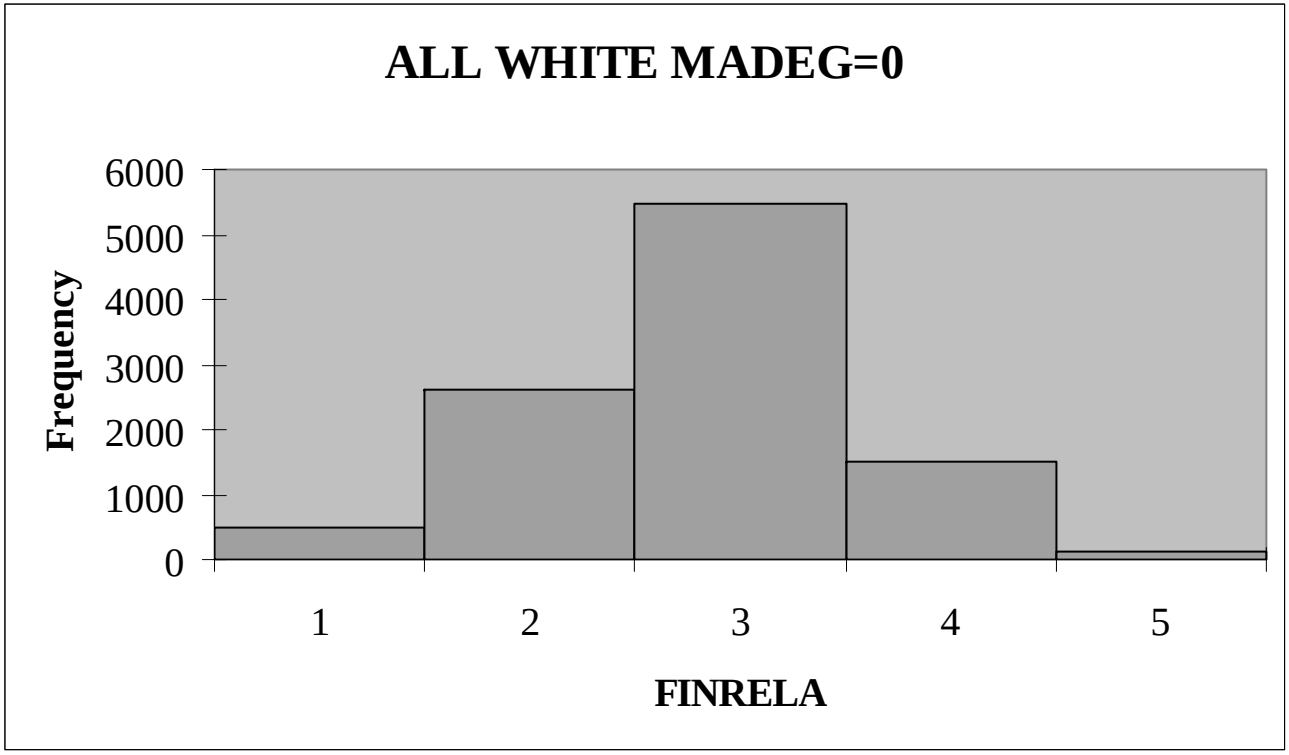
Figure 16

1977-2000 Guide Treed Regression-3



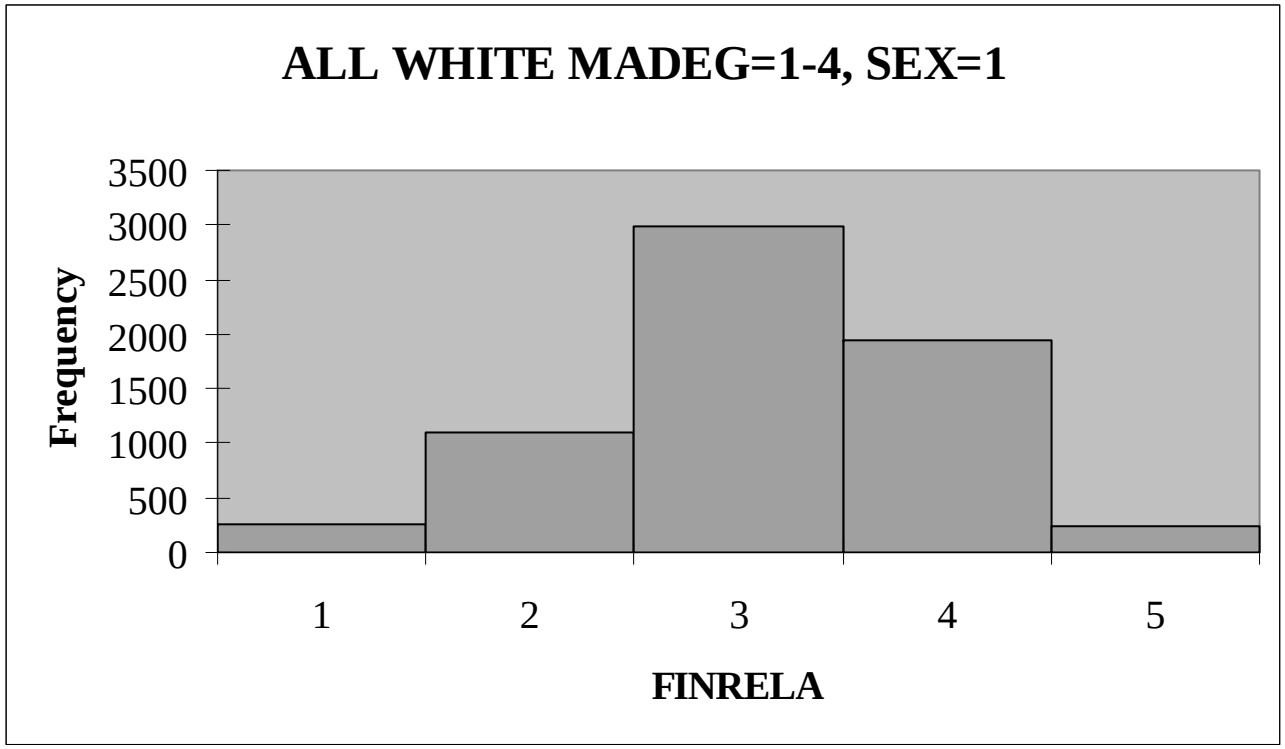
1977-2000 FINRELA Race=2,3

Figure 17



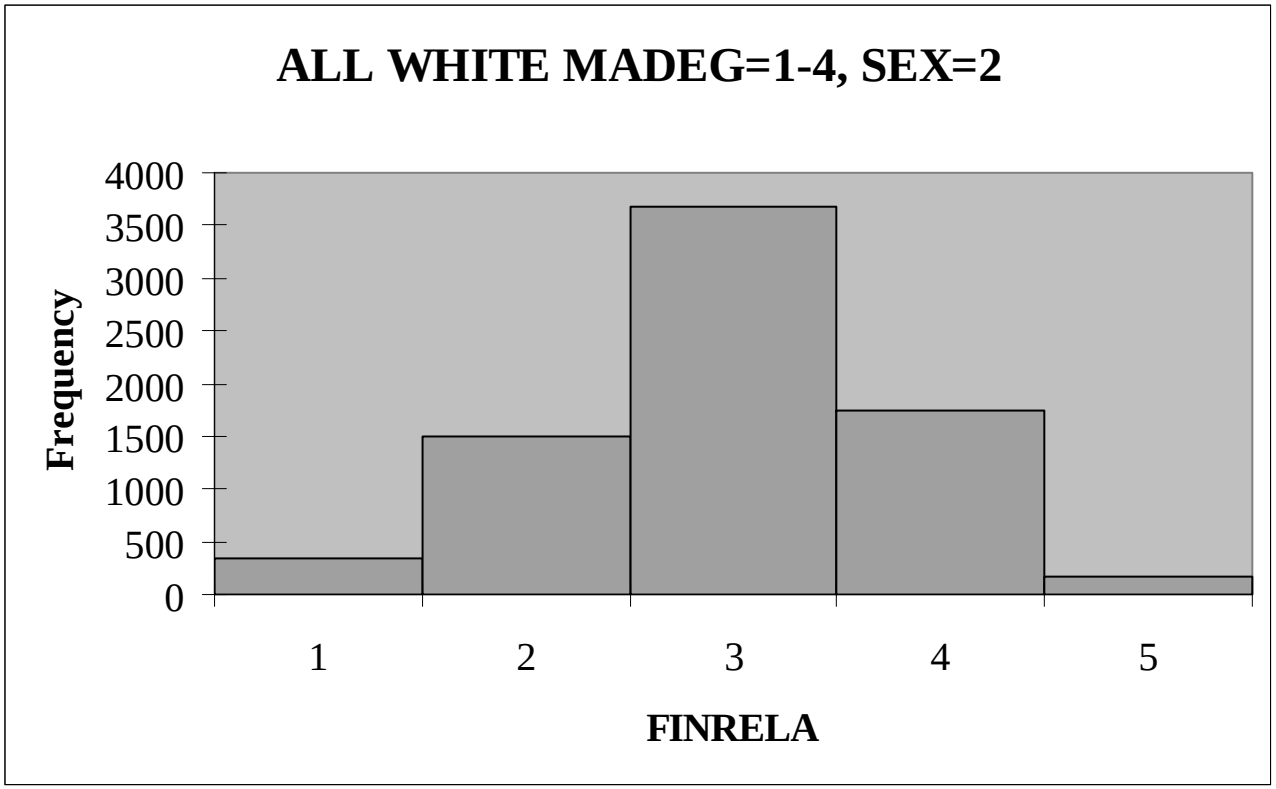
1977-2000 FINRELA Race=1; Madeg=0

Figure 18



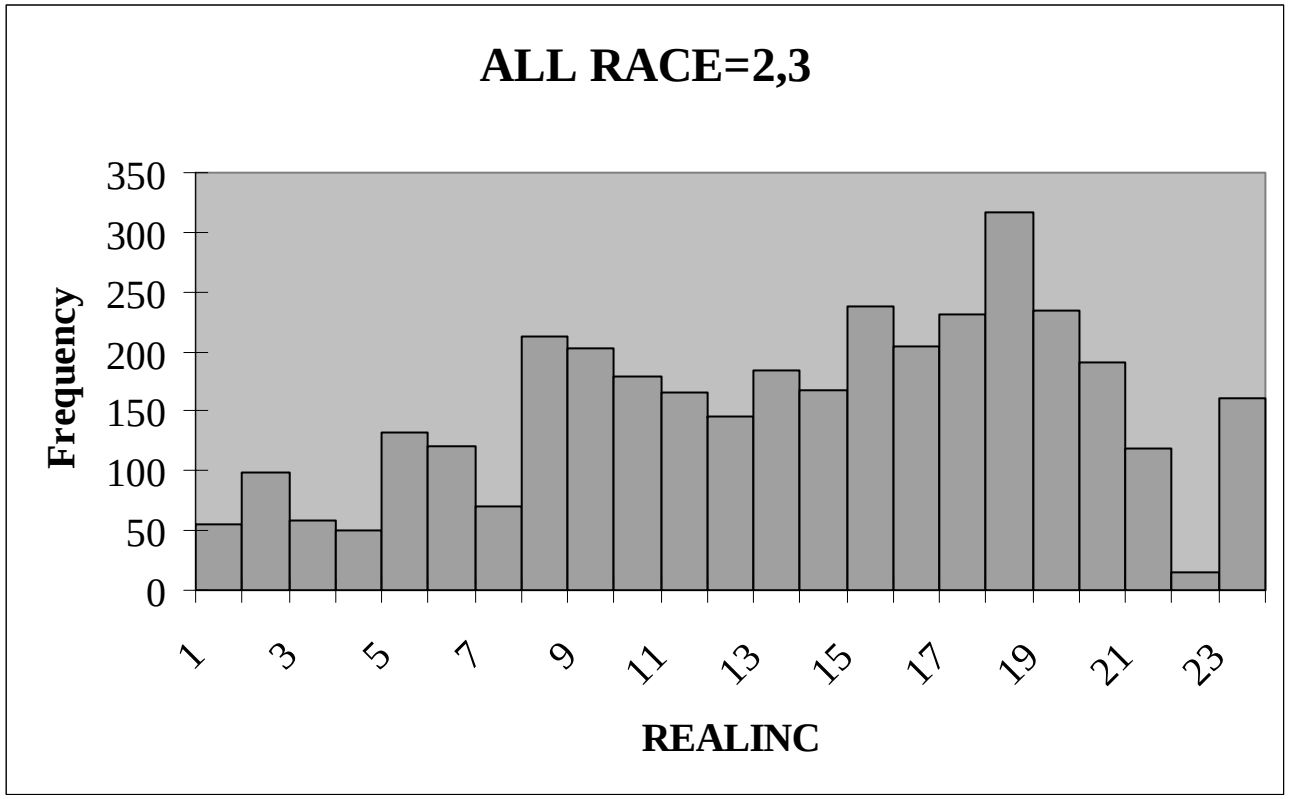
1977-2000 FINRELA Race=1; Madeg=1-4 , Sex=1

Figure 19



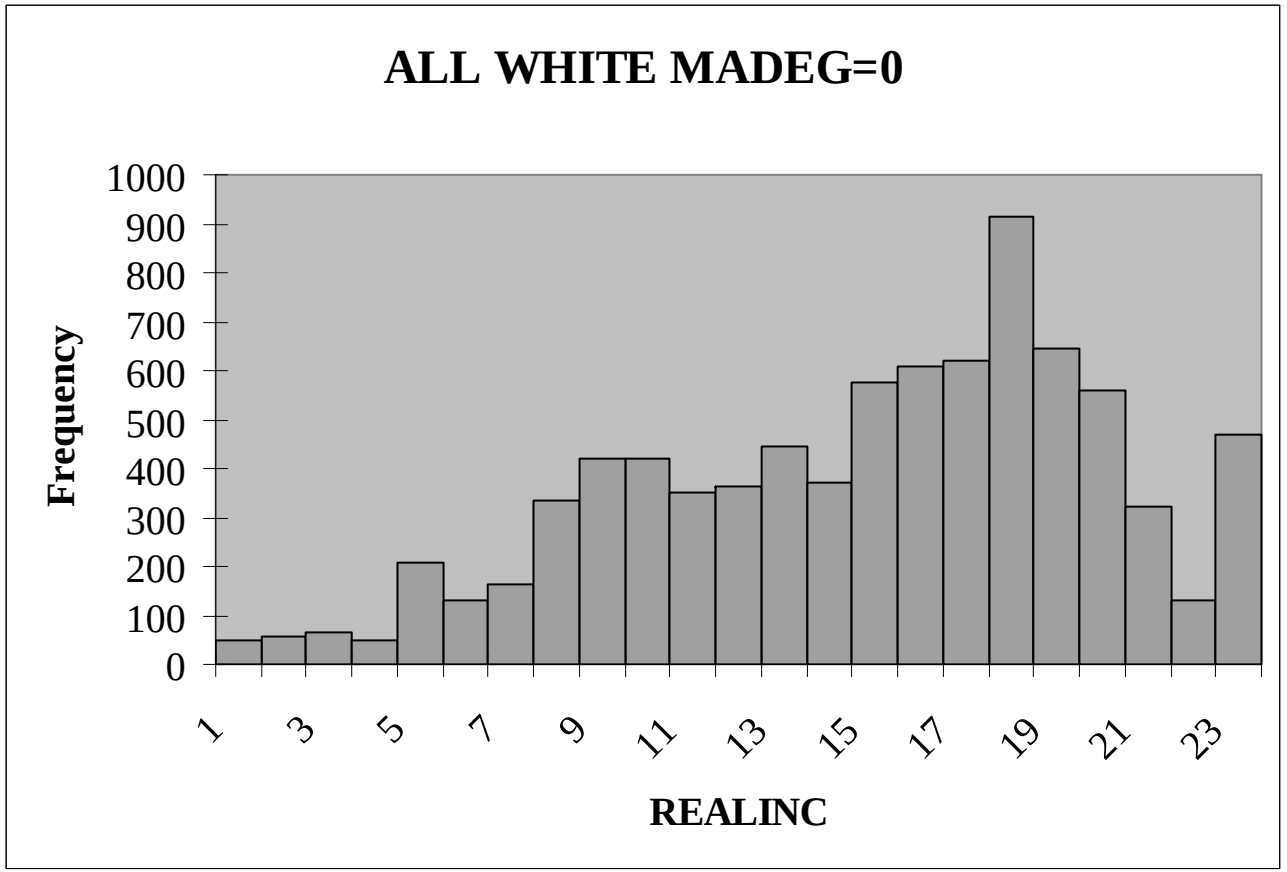
19977-2000 FINRELA Race=1; Madeg=1-4 , Sex=2

Figure 20



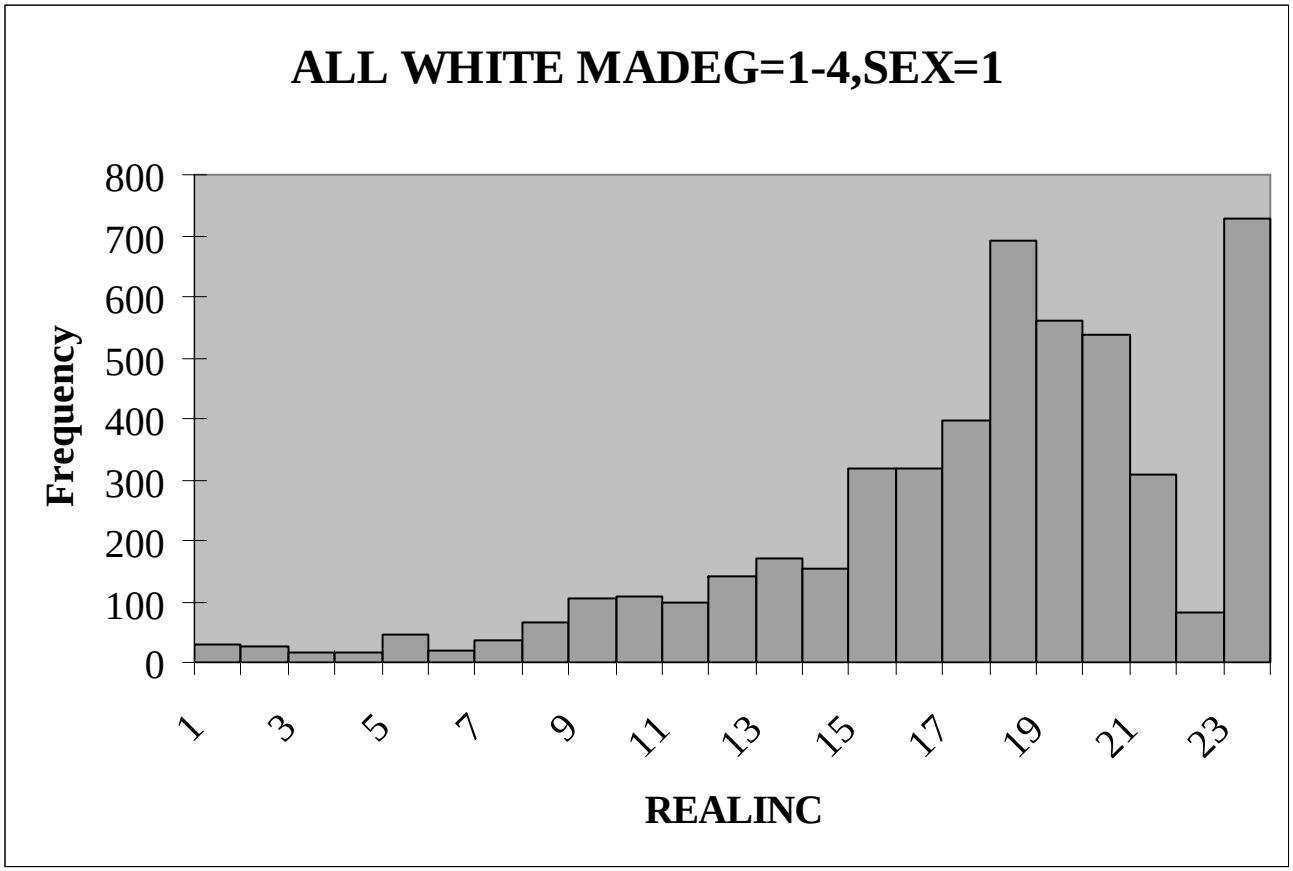
1977-2000 REALINC Race=2,3

Figure 21



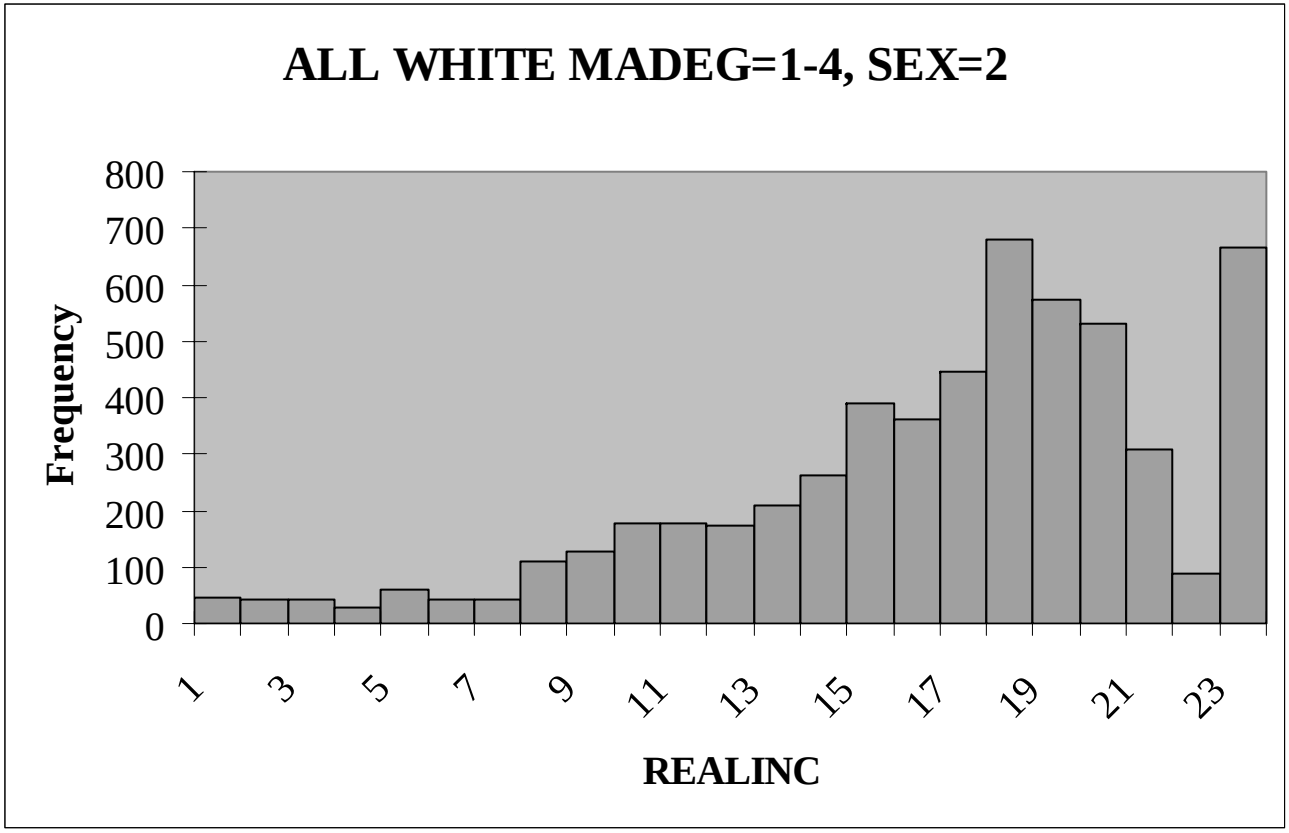
1977-2000 REALINC Race=1; Madeg=0

Figure 22



1977-2000 REALINC Race=1; Madeg=1-4 , Sex=1

Figure 23



1977-2000 REALINC Race=1; Madeg=1-4 , Sex=2

Figure 24

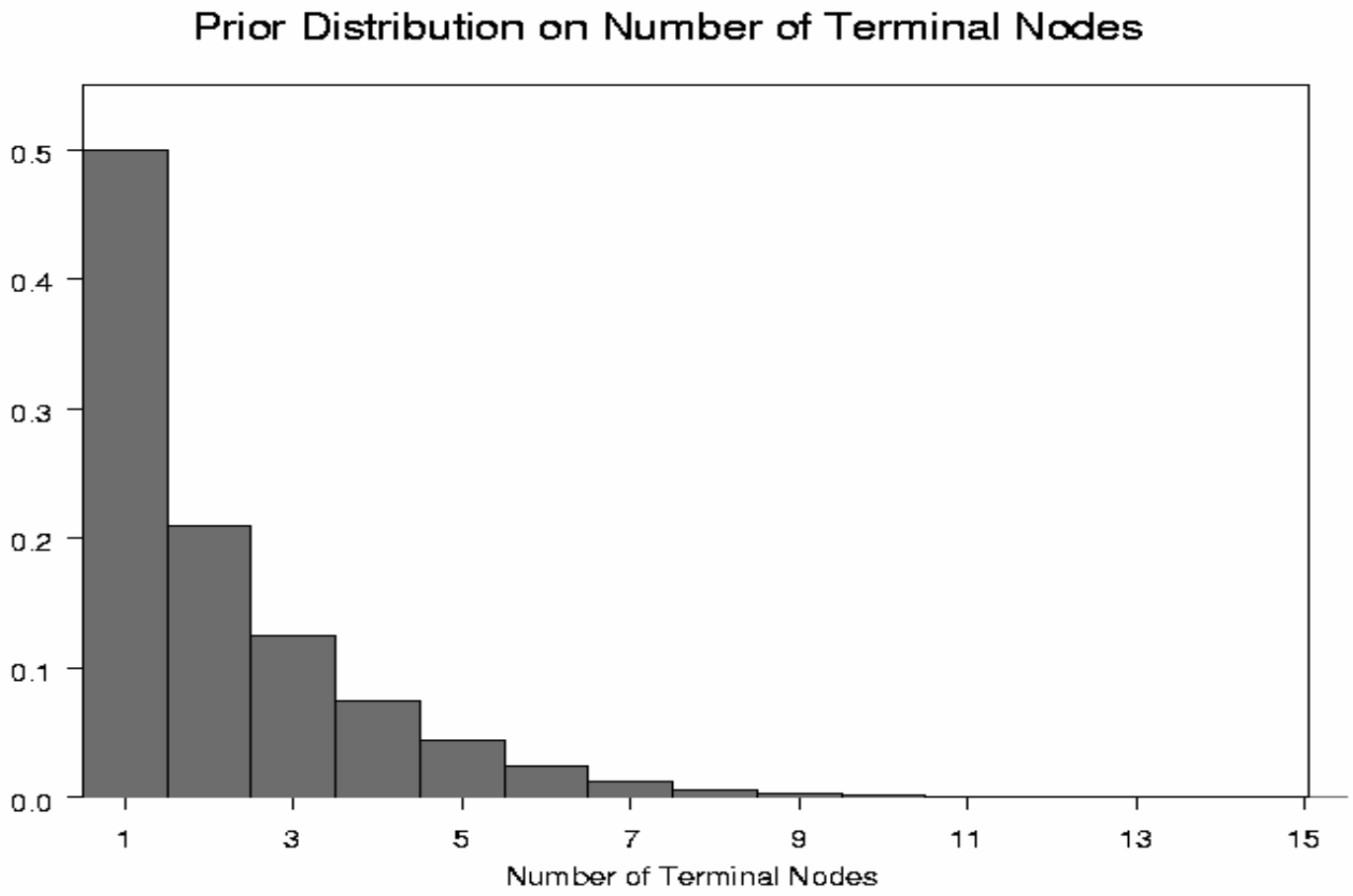

Prior Distribution on Number of Terminal Nodes
0.5
0.4
0.3
0.2
0.1
0.0
1
3
5
7
9
11
13
15
Number of Terminal Nodes

Figure 25

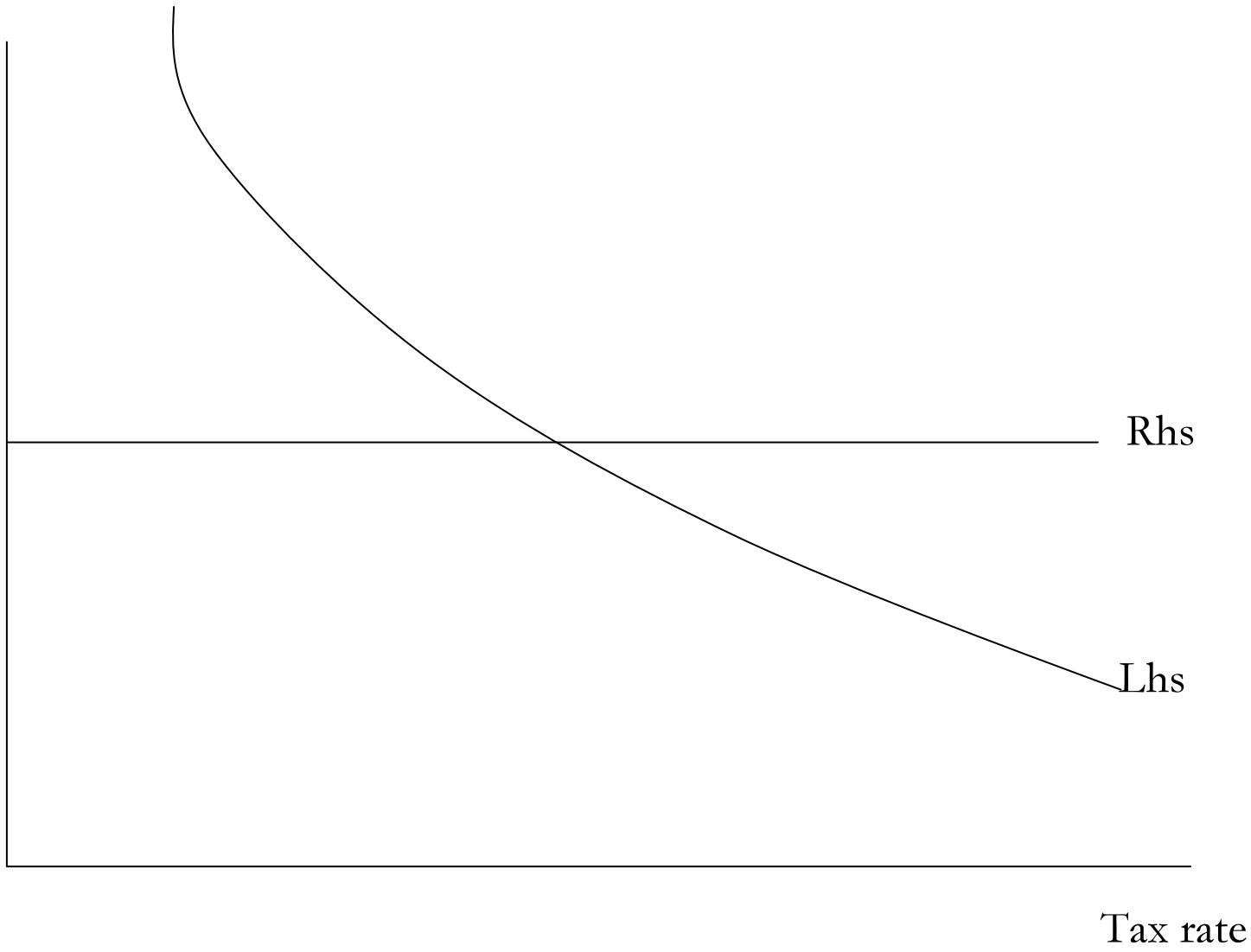


Figure 26

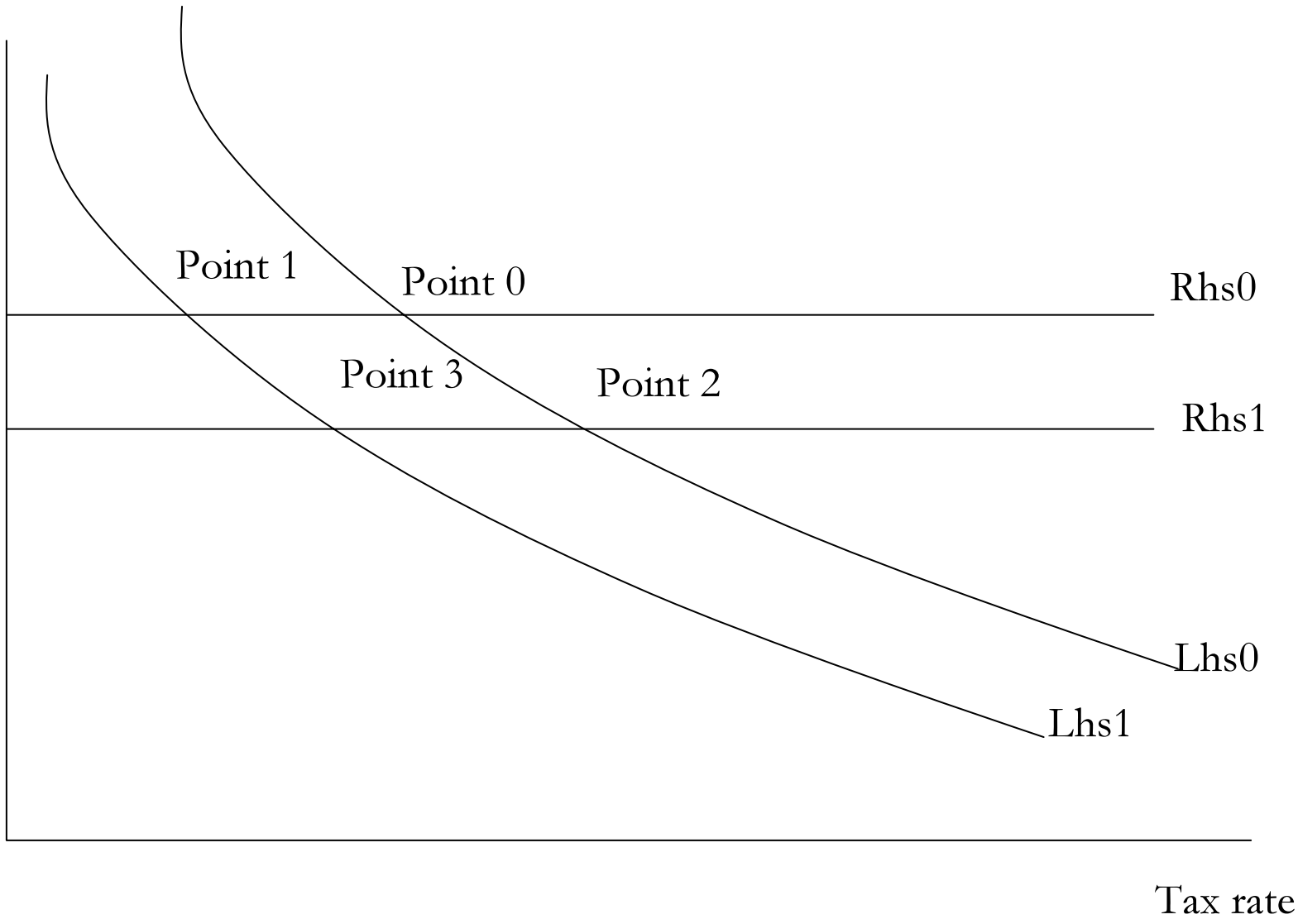


Figure 27

Tree Comparisons

C	C'	n	H(C)	H(C')	I(C,C')	maxI(C,C')	VI(C,C')	maxVI(C,C')
1994 Bayesian(Gaussian)	1994 GUIDE	1741	1.05	1.01	1.01	1.01	0.04	3.58
78-00 Bayesian (Gaussian)	78-00 Bayesian Logit	18389	1.98	1.79	1.56	1.98	0.65	4.16
78-00 Bayesian (Gaussian)	78-00 Guide	18389	1.98	1.44	1.29	1.98	0.85	4.16
78-00 Bayesian Logit	78-00 Guide	18389	1.79	1.44	1.29	1.79	0.66	3.89

Note: See Technical Appendix for variable definitions. These are comparisons of EQWLTH tree regressions.

Figure 28