Analytical Studies Branch
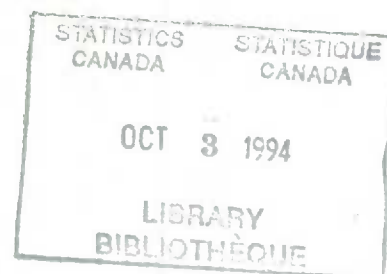
EARNINGS AND DEATH-EFFECTS
OVER A QUARTER CENTURY

by

Michael Wolfson,[1,2] Geoff Rowe,[2]
Jane F. Gentleman[2] and Monica Tomiak[2]

No. 30

1. Canadian Institute for Advanced Research, program in Population Health.
2. Analytical Studies Branch, Statistics Canada.

The analysis presented in this paper is the responsability of the authors and does not necessarily represent the view or policies of Statistics Canada.

ABSTRACT

There is widespread interest in disparities in health status across income groups and other classifications of socio-economic status. In Canada, as in many other countries, there is considerable evidence showing such disparities. This study reports an analysis of male mortality at ages 65 to 74 in relation to employment and self-employment earnings histories during the 10 to 20 years prior to age 65, as well as marital status, disability, and age at retirement. The analysis is based on administrative data from the Canada Pension Plan covering more than half a million individuals. Significant mortality gradients are found throughout the earnings spectrum. These gradients are also clearly evident in a multivariate context. The results illustrate the major potential of administrative data for research. Substantively, the results raise important questions regarding pension and health policy.

A. Introduction

There is widespread evidence that individuals who are economically or socially better off also live longer and are healthier. This contrasts with other evidence that in Canada at least, broad-based public health insurance has succeeded in providing generally equal access to medical care (Manga et al., 1987; Broyles et al., 1983).

The juxtaposition of these two lines of evidence is potentially disturbing. Why do socio-economic status gradients in health persist in a society with apparently equal access to medical care? Have we overvalued medical care as compared to other determinants of population health -- for example as discussed by McKeown (1984); or have we overstated the extent of equal access to medical care. These questions are particularly important given the large volume of resources consumed by medical care services.

Because the evidence on socio-economic gradients in health status has such important implications, studies in this area have generally been subject to intensive review and criticism. Most notable, perhaps, has been the Black Report in the U.K. (Townsend and Davidson, 1988) which charted trends in mortality rates by "social class". Social class in this report was defined in terms of the occupations reported on death certificates. A major concern is the quality and stability of the occupational coding, and the fact that some occupational groups blur together individuals with a range of socio-economic status.

This paper reports further evidence relating socio-economic and health status. More precisely, we present a longitudinal analysis of post-age 65 male mortality in relation to employment income over the previous ten to twenty years. The underlying data are of high quality; and the results that emerge are clear -- higher earnings for males in late middle age are associated with significantly lower mortality two decades later.

The plan of the paper is first to review briefly some of the main studies that have examined mortality gradients in relation to socio-economic status variables. Then several basic sets of results are presented graphically. The paper concludes with discussion of a multivariate statistical analysis.

B. Background

There are a number of questions that arise in considering associations between health and socio-economic status, particularly income or social class. One major question is the magnitude and shape of the relationship. Another much more difficult question concerns causal pathways. If higher income individuals live longer, is it because they are healthier to start, or does higher income itself predispose individuals to both better health and greater longevity? Or is the causal story far more complex with a wider variety of important factors that determine mortality as well as each other, in a way that changes with age and over time?

In principle, the only way to address these latter questions is by a careful experiment which by its very nature would be both practically and ethically infeasible. Thus, only indirect and weaker methodologies are available. To begin, it is helpful to review briefly the major kinds of evidence regarding the correlation of health status and socio-economic status (SES). More extensive reviews are given in D'Arcy (1989), and Blaxter (1986) for the U.K.

The Black Report in the U.K., as noted above, aroused a great deal of interest with its conclusion that disparities in mortality rates by social class were considerable, and that they were widening over time. One concern expressed about these results is the reliability of the occupation variable used to define social class and hence the measure of SES. Another concern is the grouped nature of the data; mortality rates are computed by age range, sex, and one of five social classes. No individual data are used so that there could be considerable heterogeneity within each of these groups. Finally, the data are a sequence of cross-sectional snapshots of the British population. Thus, while correlations may be clearly evident, it is not possible to draw inferences about whether low social class leads to higher mortality, for example, or poor health leads to both.

More detailed results for the U.K. have been derived from a longitudinal follow-up of one percent samples from both the 1971 and 1981 census (Fox et al., 1985; Goldblatt, 1989). These data do not suffer from the methodological limitations just noted, other than concerns regarding the use of "social class" defined in terms of occupation. These results also show significant and widening mortality differences.

Data on mortality rates by occupation in France (Isnard, 1989) based on longitudinal follow-up to the census suggest significant differences by social class. In contrast, Lundberg (1986) reports that Sweden has comparatively smaller gradients in morbidity than the U.K., and no clear gradient in mortality.

The first major study in the U.S. to consider individual level correlations of mortality with income and other SES variables was Kitigawa and Hauser (1973). Their results were based on the 1960 census and matched death certificates for the four months immediately following the month of the census. Kitigawa and Hauser found higher incomes associated with lower mortality among the non-elderly, but not among those over age 65. Again, these data are essentially a cross-sectional snapshot.

An update of the Kitigawa and Hauser study has recently been published by the U.S. Department of Health and Human Services (Rogot et al., 1988). These data are based on a two year mortality follow-up for almost one million individuals generally representative of the U.S. population. Over 10,000 death records were found and matched, although the authors expect that about five percent of the deaths were lost due to the limitations of the matching process. The data show clear mortality gradients among white males (the relevant group for comparison with the results presented below) both by income and by educational attainment within virtually all age ranges.

In Canada, the only broad-based population studies to date are based on grouped rather than individual data. Wigle and Mao (1980) used death certificates matched to average incomes of census tracts for the 1971 census. This analysis has been updated by Wilkins et al. (1989) using 1986 census data. The two studies show clear gradients in mortality by SES category, with an apparent decline in the magnitude of the gradient over the fifteen year period. These analyses are both subject to the limitations of cross-sectional results. They suffer further from the possibility of an "ecological fallacy". In other words, because of the heterogeneity of the population within each census tract, it is possible to obtain misleading conclusions about individual-level associations from data averaged over a group. A hypothetical example in this context is given in Duleep (1986b).

In addition to these broad national studies, many studies focus on more specific populations. Most notable of these is the study of public servants in the U.K. Marmot (1986) shows a striking gradient in mortality over a ten year period by occupational grade in the civil service. This study is stronger than the Black report evidence insofar as it is longitudinal, and has a more precise variable to indicate SES, namely grade in the civil service. In addition, a large number of variables such as family background and elements of blood chemistry were ascertained at the beginning of the study, so that along with the cause of death information it is possible to begin proposing more specific causal pathways for the observed SES gradient. An obvious limitation of this evidence is that it applies only to a special sub-population.

Hirdes and Forbes (1989) report a strong association of mortality with income from the 20 year Ontario Longitudinal Study of Aging. One of the interesting results was the importance of smoking behaviour in relation to the strength of the association of income and education to mortality. In a multivariate analysis that included smoking, the association of mortality with income remained significant, but the association with education did not. Furthermore, in an earlier analysis of the same data (Hirdes et al., 1986), an effort was made to disentangle the sources of this association. They conclude that "(I)t is more likely therefore that a change in income leads to a change in health. The mechanisms involved may include stress, changes in purchasing patterns, and/or changes in social and physical activities."

The study that is most similar to this one in terms of data used is that of Duleep (1986a, 1989). She uses a special sample of Social Security administrative data records. These records have been exactly matched to death certificates in a six year follow-up period, and to a Census Bureau sample (the March Current Population Survey) giving data on other SES variables such as family income. From the Social Security data, she thus has year-by-year employment earnings for the years prior to 1972, and year-by-year survival for the years 1973 to 1978.

Duleep's analysis focuses only on white married males aged 35 to 65 (in 1972) and is restricted to a sample size of about 10,000. She finds a clear gradient in mortality with respect to income at the lower and middle ranges of the income spectrum, but very little impact of income at higher levels (above $10,000 in 1972). While those who completed college have lower expected mortality, there is no clear relationship for the other lower levels of educational attainment. These limited findings may be due to the correlation between income and educational attainment, both of which were used as independent variables in the regression analysis. If income or education alone were used as explanatory variables, it could be that a clear mortality gradient throughout the SES spectrum would have been found.

C. The Data

For this analysis, data have been drawn from the administrative records of the Canada Pension Plan (CPP). This is a public earnings-related pension covering (along with the identical Quebec Pension Plan, QPP) 100% of the Canadian paid labour force. The plans commenced in 1966. For their operation, employment earnings (both of employees and the self-employed) are subject to a payroll tax administered annually as part of the income tax system. In turn, everyone who has contributed for at least three years is eligible for a

retirement pension at age 65, and a lump sum death benefit.[1] The retirement pension depends on year-by-year employment earnings between the ages of 18 and 65 according to a complicated formula.

Given this program structure, virtually everyone who has worked in Canada outside the province of Quebec and who attains age 65 will become a beneficiary of the CPP. The year and month of death are recorded on the administrative data file both for purposes of terminating retirement pension benefits and to pay the lump sum death benefit. Revenue Canada Taxation is the source of the employment income numbers while individuals are contributing to the CPP. The CPP file also contains earnings data from the QPP so there are no missing earnings data for CPP beneficiaries who spent parts of their working careers in the province of Quebec. Thus, both the date of death and the year-by-year earnings history variables are considered to be of high quality.

The CPP beneficiary file that has been used contains over 5 million records. As a first stage in the analysis of these data, the focus is on males. (Future analysis will also consider females.) The analysis is restricted to those males who attained age 65 on or after September 1, 1979 (545,769 individuals). This date was chosen for two reasons. First, the CPP/QPP had been in existence for over a decade, so the take-up rate for retirement pensions was virtually 100%. Second, this assured at least 13 years of year by year earnings history prior to attaining age 65 for all the observations used. Just over 10% of this population (55,101) had died by September 30, 1988 -- nine years and a month later -- the cut-off point used when the data were extracted from the administrative file in the Spring of 1989.

The CPP population generally comprises the majority of the relevant population. For example, 87 percent of all males age 55 to 59 in 1986 are recorded in Revenue Canada's personal tax return files as having contributed to the CPP or QPP at some point in the last 20 years. Another 2% contributed but did not file tax returns. Most of those who were not CPP contributors had very low incomes.

Table 1 gives more details for males age 55 to 59 based on data from personal income tax returns for 1986 and from the 1986 census. The estimated 549,000 male tax filers in this age range were divided into percentile groups as shown in the first column. The second column shows the maximum total income of the tax filers in each group.

The third column shows the proportion that had ever contributed to the CPP. This proportion is over 95% except in the bottom fifth, and is 61.5% for the bottom ten percent. This corresponds to the fact that many of those in the lowest income ranges are receiving income from sources other than earnings or self-employment -- for example bond and bank interest, dividends, and private pensions.

As well, 25 thousand male tax filers in this age group reported negative self-employment income, losses which are not counted as earnings subject to contributions, but do offset income from other sources in determining total income. Many of these tax filers therefore appear to be in the lower total income ranges, though the ability to incur such losses is probably indicative of substantial wealth (e.g. collateral assets) rather than poverty.

---

[1] There is also a provision for flexible retirement age which has been introduced recently. However, this does not affect the analysis and has been ignored.

This is also evident in the fourth column which shows the percentage earnings for C/QPP purposes are of total income. This exceeds 100% in the bottom decile, in part because positive employment income is being offset by negative self-employment income, and because of tax shelter and other losses. Otherwise, earnings which are taken into account in the CPP data generally amount to over 80% of total income.

The last column compares the tax filer data with the 1986 Census. The tax filer decile groups each contain about 55,000 individuals. The census data show almost twice as many individuals below the dollar cut-off for the bottom decile of tax filers. This is as expected because many very low income individuals do not need to file tax returns. Other than in the bottom decile, the figures from the census are very close to the tax data. (Note that the expected number in the 90-95 and 95-100 percent groups is about 27,500.)

Table 1: Distribution of Male Taxfilers Aged 55-59 in 1986 by Total Income

Total Income Percentile Group	Maximum Total Income	Percentage Ever Contributing to C/QPP	Percent of Total Income that is Earnings	Counts from 1986 Census in Same Total Income Ranges
0-10	6,032	61.5	128.5	100,370
10-20	12,165	89.7	56.4	53,345
20-30	17,942	95.8	66.1	60,850
30-40	22,594	96.6	73.2	54,855
40-50	27,060	98.6	79.2	60,440
50-60	31,376	99.0	83.1	56,210
60-70	36,278	99.0	86.1	51,175
70-80	42,606	99.3	87.5	50,795
80-90	54,598	98.8	87.3	52,865
90-95	70,228	99.4	85.8	26,615
95 +		99.1	82.3	27,375

Source: Special tabulations of 3% 1986 taxfiler sample and from 1986 Census.

Note: The table covers 549,488 tax filers. 7,904 tax returns of decedents, emigrants, and those claiming disability were excluded. Among these excluded returns, 5,614 had C/QPP contributions. The census data show a total of 594,895 males in this age range.

D. Initial Results -- Earnings and Death

Figure 1 presents the overall relationship between earnings histories and mortality. The horizontal axis shows employment income averaged over each individual's career from 1966 to the year before he reached age 65[2]. Earnings are in 1988 dollars, and each person's annual earnings were "updated" or re-scaled using the average industrial wage index before the average was computed.

(It may be noted that this resulting "updated average earnings" figure, since the updating is based on a wage index rather than the Consumer Price Index, might be more appropriately interpreted as an index of relative position on the earnings spectrum. Also, as just noted in Table 1, it fails to take account of other sources of income such as government transfers and investment returns, and the incomes of other family members.)

---

[2] The year the individual reached age 65 is excluded because, had he worked at all that year, he would likely have worked for only a fraction of the year; see Kennedy (1987). Also, all years after the last year of non-zero earnings (i.e. trailing zeros in the earnings vector) have been excluded from the average.

Mortality Rates: Ages 65-70

For 11 Percentile Earnings Groups [see text]

3%

5%

10%

20%

20%

20%

10%

5%

3%

2%

2%

95% Confidence Intervals − − − − − −

Mortality Rate

0.20 0.15 0.10 0.05 0.00

Average Earnings: Ages 43-64
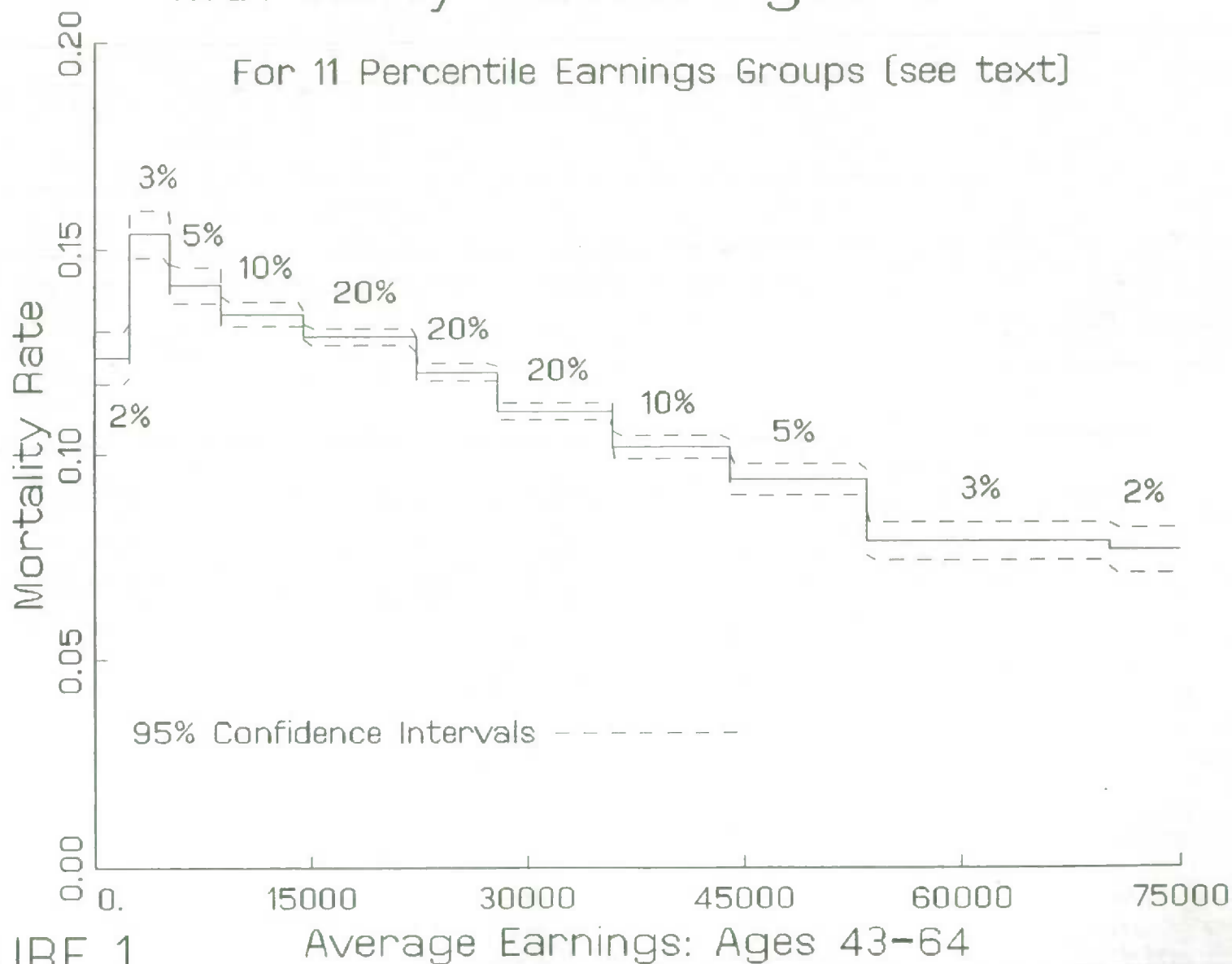
0. 15000 30000 45000 60000 75000

FIGURE 1

The vertical axis shows the probability of dying over the five year period from exact age 65 to exact age 70, conditional on reaching age 65. In fact, data are available on deaths up to age 74 for a subset of the population. The age 65 to 70 interval was chosen because the coverage of the data is greater (fewer observations are right-censored, i.e. affected by the September 1988 cut-off date), and it is a convenient interval for comparison with other studies.

In order to compute the five year mortality rates for the different earnings ranges, the males in the population were first sorted in increasing order of their average updated earnings, and then grouped based on percentiles. A total of 11 groups were defined, dividing the population at the 2nd, 5th, 10th, 20th, 40th, 60th, 80th, 90th, 95th, and 98th percentiles. The vertical steps in the curve of Figure 1 thus correspond to the average earnings cut-points for these percentiles. For example, the top two percent of the population (almost 11 thousand observations) had average updated earnings above about $70,000.[3] The five year mortality rates were then computed for each percentile earnings group,[4] along with a 95 percent confidence interval.[5]

Aside from the lowest earnings group, there is a clear monotone and statistically significant pattern -- higher income males experienced lower mortality. The "blip" at the bottom of the earnings range likely reflects the fact that individuals with close to zero earnings over one to two decades of their lives probably depended on other sources of funds such as government transfers, investments or other family members, and thus may not have had incomes as low as those recorded from earnings alone.

It is difficult to imagine a clearer and more unequivocal result. These data cover over half a million individuals, and for each individual data from almost a quarter century of their lives have been drawn upon.

It should be emphasized that these are not cross-sectional results. The earnings shown along the horizontal axis were received between the ages of 43 and 64 -- on average 10 to 20 years before the mortality experience being considered. This point is illustrated in Figure 2 which highlights the longitudinal character of the data being used.

These results are also notable in that they show the mortality gradient continuing beyond the lower-middle income ranges -- unlike Duleep (1989) for example, and a gradient amongst the elderly -- unlike Kitigawa and Hauser (1973). These results are however similar to the findings of Wilkins et al. (1989) for Canada, Rogot et al. (1988) for the U.S., and the Black Report for the U.K. (Townsend and Davidson, 1988). One caveat should be borne in mind regarding this similarity: the last three studies are all essentially cross-sectional; the measure of SES is either contemporaneous or within one or two years of the measure of mortality. In this analysis of the CPP data, in contrast, the mortality experience follows the SES measure (earnings) by decades.

---

[3] The dollar cut-points are as follows: 2% - $2,404, 5% - $5,137, 10% - $8,745, 20% - $14,494, 40% - $22,279, 60% - $27,991, 80% - $35,987, 90% - $44,049, 95% - $53,500, 98% - $70,069.

[4] These mortality rates were calculated using the product limit form of estimator which takes censoring into account.

[5] This confidence interval is based on the assumption of homogeneity within each earnings group. Frequencies of death are assumed to be conditionally binomial. Since there is an apparently continuous gradient of mortality with earnings, the homogeneity assumption is an approximation, so that the range of the confidence interval is a lower bound.

While Figure 1 shows five year mortality from age 65 to 70, the data can provide mortality rates for nine years. Those males who became 65 during September 1979 would be 74 by September 1988, the last month of data. Such individuals correspond to the top diagonal line in Figure 2, and would have been age 52 in 1966.

Survival probabilities by month after age 65 to age 74 are shown in Figure 3, conditional on reaching age 65. This time the survival curves are for five earnings quintile groups rather than the 11 percentile groups of Figure 1. The mortality gradient with average pre-retirement earnings continues throughout the nine year period -- the curves do not cross and the distances between them gradually become wider.

The gradient in Figure 1 is related to the curves in Figure 3. If Figure 1 had used only cut-off points at the 20th, 40th, 60th, and 80th percentiles, then the mortality rates measured along the vertical axis would correspond to one minus the survival probabilities along a vertical line at age 70 in Figure 3. Figure 1 in comparison to Figure 3 thus gives more detail at either end of the earnings spectrum, but only for a single age. Figure 3, in contrast, shows the robustness of the mortality gradient over almost a decade of follow-up, but for fewer earnings groups.

Figure 3 also shows overall male survival probabilities for Canada in 1985-87, centered on the 1986 census. These overall data show higher average mortality (i.e. lower survival probabilities with the dashed line between the first and second earnings quintile survival curves).

The main explanation for this difference is that the CPP data exclude those with no employment income. Table 1 suggests they are primarily the poor living on government transfers rather than the very rich living exclusively on investment income. Thus it is consistent with the results so far that if the CPP data exclude a group with generally lower average incomes, then CPP beneficiaries should have higher survival rates than the general population.[6]

E. Initial Interpretations

One major question in interpreting this gradient of mortality in relation to earnings is the role played by illness. One plausible hypothesis is that a chronic illness sets in which then leads both to lower earnings and to increased mortality. Figure 4 casts doubt on this interpretation, however, at least for a significant sub-population. It is exactly the same as Figure 3 except that only a subset of CPP contributors has been considered -- those whose earnings were generally increasing year after year prior to retirement.[7]

---

[6] Another factor that could account for higher survival rates among CPP beneficiaries is the 1986 census undercount, estimated to be about three percent. This magnitude of undercount could depress male survival probabilities from age 65 to 75 by 0.4 percentage points. Also, the CPP data generally exclude residents of Quebec which has about one-quarter of Canada's population. The Quebec survival rate for males age 65 to 75 was about four percentage points lower than the rest-of-Canada rates for 1981 and 1986 (68.3% and 70.2% for all Canada except Quebec versus 64.8% and 65.7% for Quebec in 1981 and 1986 respectively).

[7] More precisely, only the 103,741 observations among CPP contributors who had a statistically significant (at the five percent level) positive rank order correlation of age and earnings were considered.
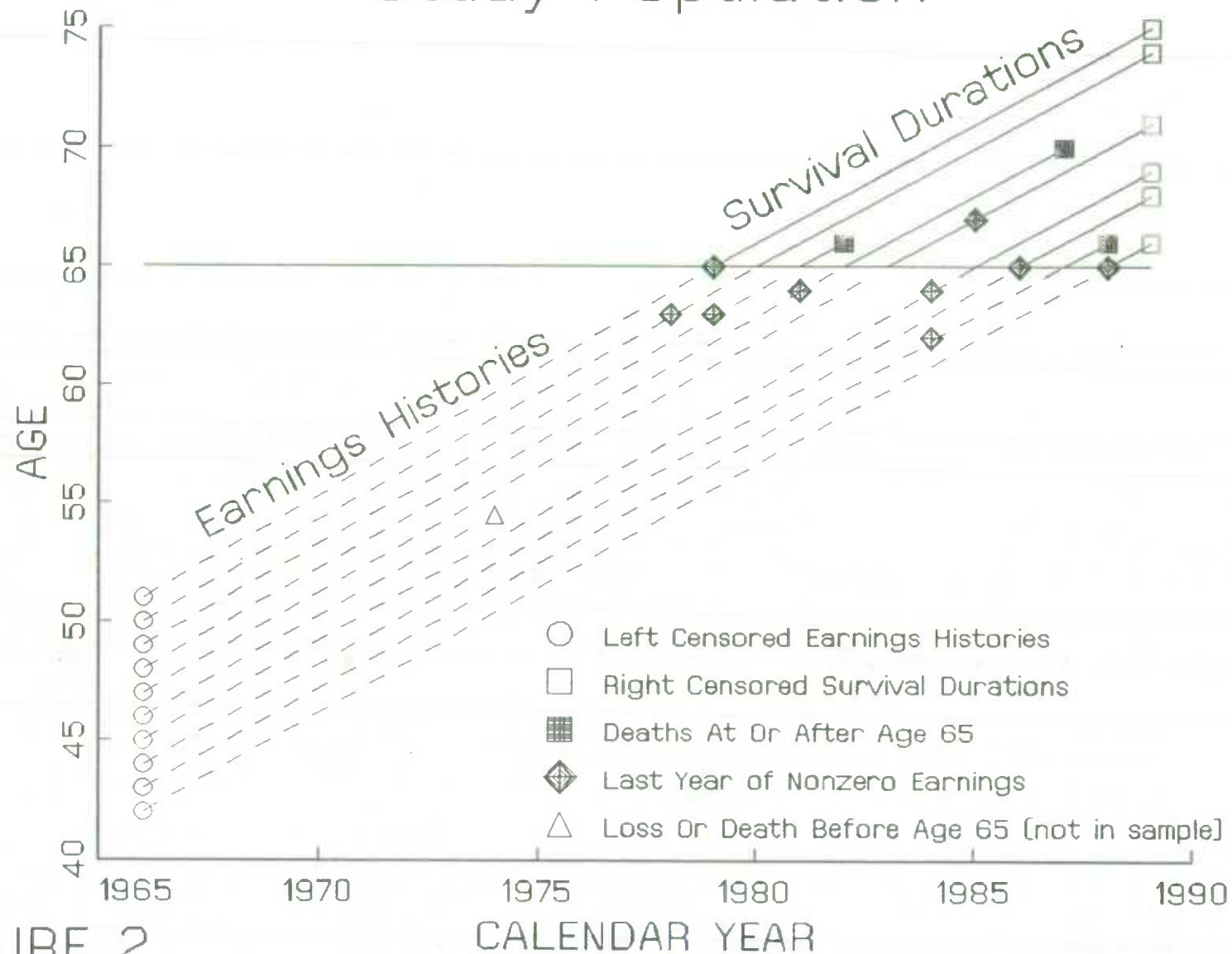
FIGURE 2

Study Population

Survival Durations

Earnings Histories

AGE

CALENDAR YEAR

○ Left Censored Earnings Histories
□ Right Censored Survival Durations
▦ Deaths At Or After Age 65
◈ Last Year of Nonzero Earnings
△ Loss Or Death Before Age 65 (not in sample)

Male Survival Curves
Conditional on Survival to Age 65

Cross-section at Age 70

Earnings Quintile Groups

#1   Earnings<= $9135

#2 $9135 <Earnings<= $16990

#3 $16990 <Earnings<= $23438

#4 $23438 <Earnings<= $31585

#5 $31585 <Earnings

⊙ Canada 1985-87

#5
#4
#3
#2
⊙
#1

Probability

Ages (Years & Months)
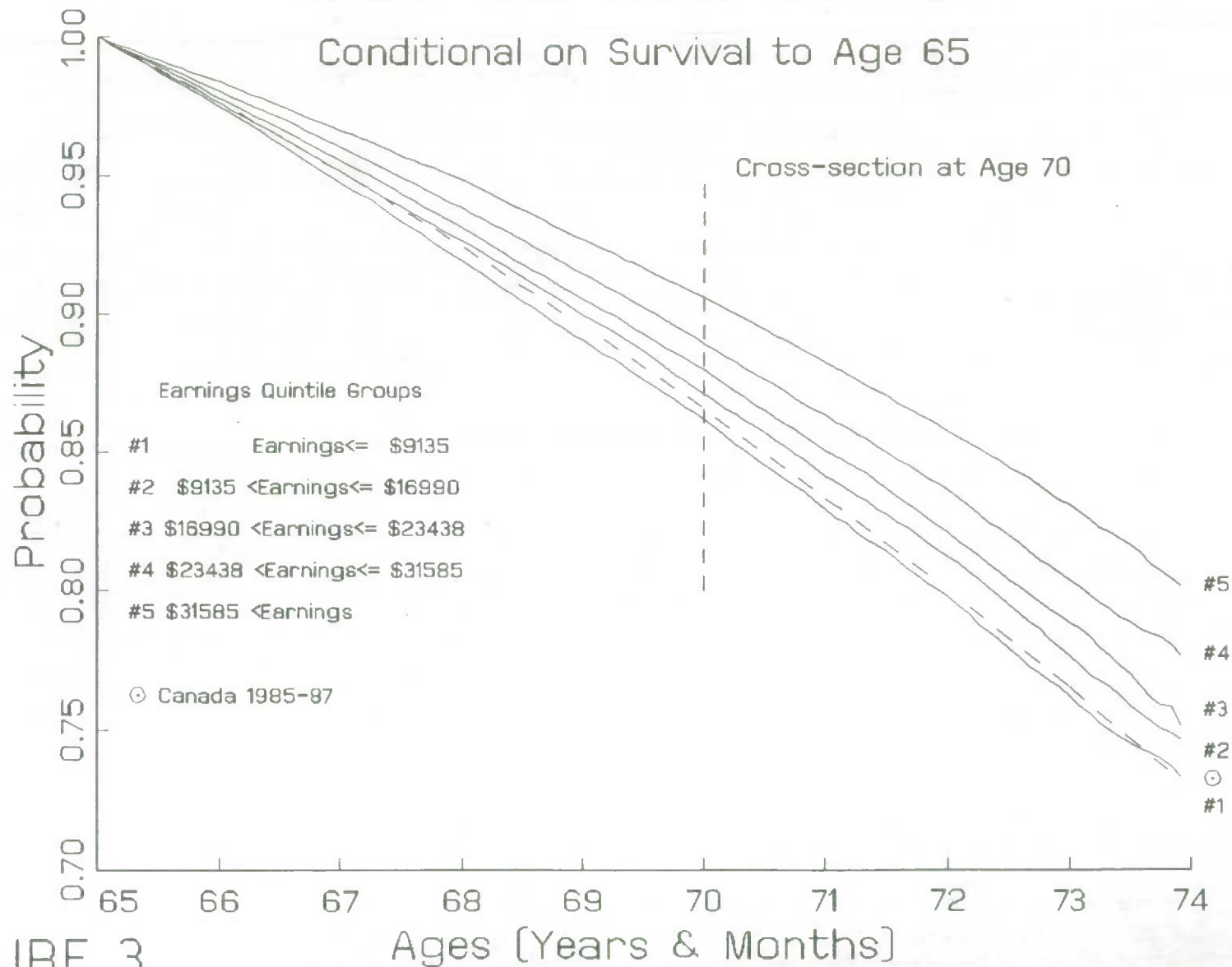
FIGURE 3

Male Survival Curves

Conditional on Survival to Age 65

For Earnings Quintile Groups

Sub-Sample With Increasing Earnings

FIGURE 4

Ages (Years & Months)
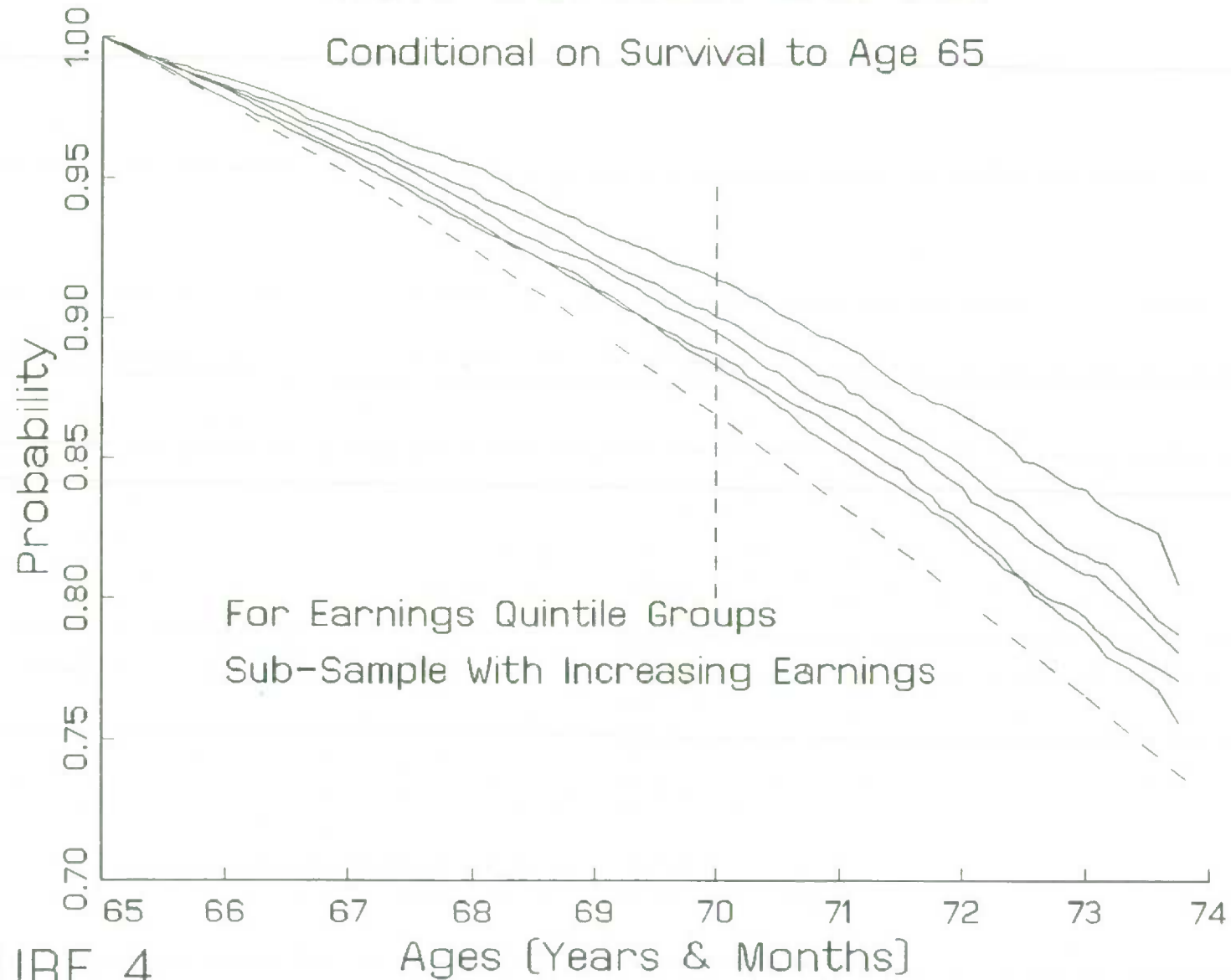
Probability

Except among the first two earnings quintiles, there is a clear and consistent increase in survival probabilities with earnings at all ages in the interval. It is difficult to reconcile these data with the hypothesis that illnesses fully account for the gradient in mortality. It does not seem plausible that these individuals became ill in their forties and fifties, and thus predisposed to higher mortality after age 65, yet continued to work and earn incomes that grew at a higher rate than average wages from their forties and fifties up to age 65. (Recall that earnings have already been deflated by the average industrial wage index.)

In order to give an indication of the importance of the mortality gradients with earnings shown in Figures 1 and 3, a simple comparison can be made with the results from cause-deleted life table analysis. Nagnur and Nagrodski (1987) estimate with 1981 all-Canada mortality rates that survival probabilities to age 75 for males who have already survived to age 65 would increase, for example, by about eight percentage points if cancer as a cause of death were eliminated (and mortality rates from all other causes of death were unchanged). The data underlying Figure 3 suggest an almost identical improvement in survival probabilities from age 65 to 75 if the CPP cohort had all experienced the mortality rates of the top 20% of average earners rather than their observed mortality rates.

In other words, the elimination of cancer would have roughly the same impact on mortality for this group as elimination of the mortality gradient by earnings for the bottom 80 percent. This is ironic given the much stronger connection in the public's mind between cancer and health, and the much larger research and medical care expenditures on cancer than on the connections between SES and mortality.

F. Further Results -- Disability, Marriage, and Retirement

Published statistics (e.g. Statistics Canada, 1980) have for a long time shown that married men have lower mortality than their single counterparts. Figure 5 shows the relationship of mortality both to earnings and marital status using the CPP data.

The horizontal axis in Figure 4 is the same as Figure 1. However the vertical axis shows survival probabilities -- one minus the mortality rates shown in Figure 1 -- conditional on attaining age 65.

The 545,769 males in the CPP data set have first been divided into three groups: (1) those who have ever received a disability benefit from the CPP ("disabled" based on the very strict definition of disability used in the CPP; 49,610 observations; 7,062 deaths between ages 65 and 70); (2) those who were married at age 65 and were not "disabled" (411,115 observations; 27,004 deaths); and (3) those who were neither "married" nor "disabled" (80,829 observations; 8,802 deaths).

Then, within each of these three demographic groups, individuals were sorted in increasing order of their average updated earnings, and divided into the same percentile groups as used for Figure 1. Note as a result that the 98th percentile cut-point for the "disabled", for example, is at about $49,000 while the same cut-points for the "married" and "not married" are at about $74,000 and $57,000 respectively.

The "disabled" have the highest mortality (i.e. lowest survival) rates, and generally lower earnings. As expected, married males have lower mortality than their unmarried counterparts -- given they are not "disabled".

The mortality gradient with earnings is again evident among the non-disabled within each marital status group, though it is not as steep. In Figure 1, mortality rates for the top 10% of the population are about half those of the bottom 10%. In Figure 4, in contrast, mortality rates of the top relative to the bottom 10% of married and not married men are about three-quarters and two-thirds respectively. Also, a "blip" in the lowest earnings range is again evident as in Figure 1.

Figure 6 shows another disaggregation of the data. Given the thirteen or more years of earnings data for each observation, an interesting question is the role of other attributes of these earnings streams or vectors, over and above the (updated) average earnings that has been examined so far.

One such attribute is the last year with non-zero earnings, which we interpret as the year of retirement whether an individual worked to age 65 or retired early. While the date of retirement is not available directly from the CPP data on earnings histories, the last year with non-zero earnings can be readily observed. If this was age 61 or before (and there was no "disability"), the person is considered an "early retirement" in Figure 6 (115,771 observations; 7,520 deaths). Those who were not "disabled" or "early retirements" were then divided into two groups -- those who had non-zero earnings in every year until retirement ("uninterrupted work history"; 279,023 observations; 20,910 deaths), and those with at least one year with zero earnings prior to their last year of earnings ("interrupted work history"; 97,150 observations; 7,376 deaths). Again, each of these four groups was further subdivided by average updated earnings percentiles as in Figure 5.

Figure 6 shows no significant differences in mortality among late retirees between interrupted and uninterrupted work histories. Both show some gradient with earnings. However, there is a sharper difference between early and late retirees. Early retirees generally have higher mortality, and a steeper gradient with earnings.

This latter phenomenon might be explained by greater heterogeneity among the early retiree population. Those at the lower end of the earnings spectrum might be workers laid off in their late 50s unable to find another job, or workers who had to quit work due to their deteriorating health (but not so ill as to qualify for a disability benefit under the CPP), or the deteriorating health of a spouse. Those at the upper range of earnings, on the other hand, might have been so well off both financially and in terms of their health that they decided to retire early in order to enjoy themselves.

G. Multivariate Analysis

The results so far suggest that a variety of factors are importantly associated with mortality, including average earnings, whether disability benefits have ever been claimed, marital status, and various attributes of individuals' pre-retirement earnings histories such as the age at which earnings ceased. Thus some form of multivariate analysis appears warranted.

It is clear from the results in Figures 5 and 6 that all of these groups of factors are significant. Moreover, the figures suggest that the interactions may not be sufficiently well-behaved to be represented in a single equation regression analysis, even with many interaction terms on the right hand side. For example, it does not seem appropriate a priori to impose an assumption of constant proportional hazards for marital status or age at retirement.

Survival from Age 65 to 70

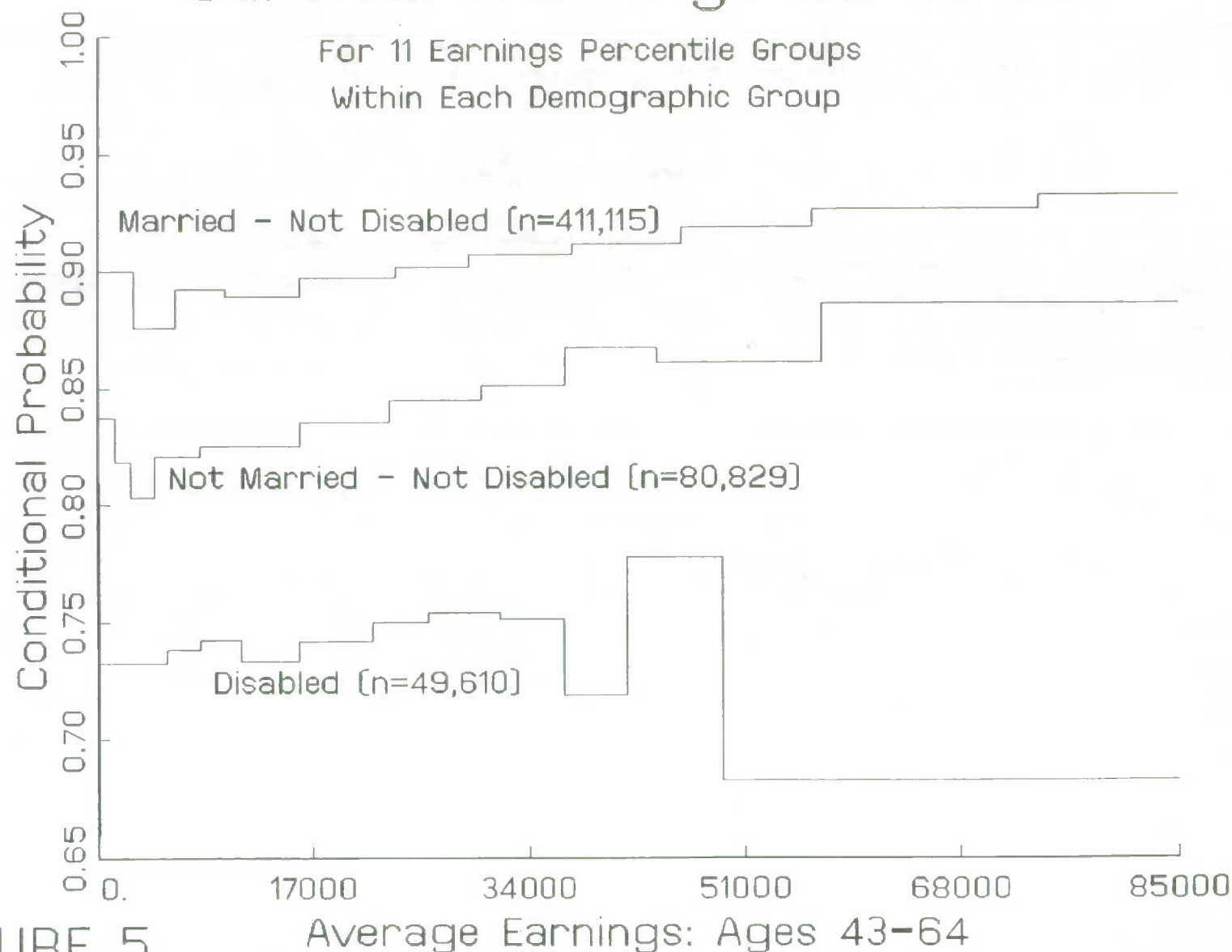For 11 Earnings Percentile Groups
Within Each Demographic Group

Married - Not Disabled [n=411,115]

Not Married - Not Disabled [n=80,829]

Disabled [n=49,610]

Conditional Probability

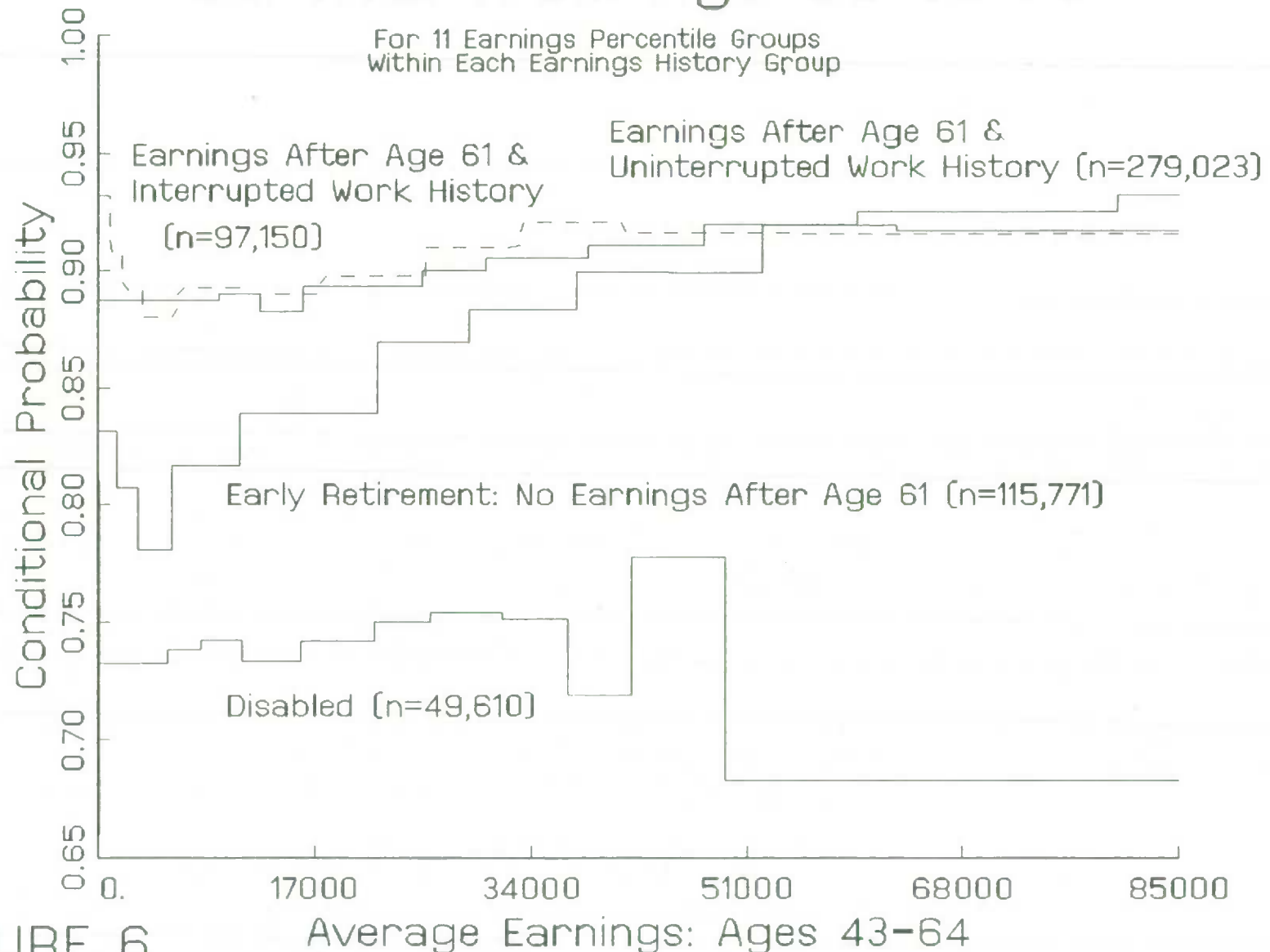Average Earnings: Ages 43-64

FIGURE 5

FIGURE 6

Thus, a set of independent regressions have been estimated, one for each of two marital states and 13 distinct ages at retirement, for a total of 26 regressions. This specification is the culmination of considerable analysis, and is intended to present the key multivariate results in as clear and parsimonious a manner as possible.

The regression specifications are linear models of the form $y(i) = x b + s e(i)$, where $y(i)$ is the natural logarithm of $t(i)$, the number of years lived beyond age 65; x is a row vector of covariates; b is a column vector of unknown coefficients; and i indexes individuals. In least squares multiple regression, the error term $e(i)$ is from a standard Normal distribution, and s is a constant scale parameter. Here, as is commonly done for lifetime data, we assume instead that $e(i)$ is from a standard extreme value distribution. This is equivalent to a proportional hazards model with a Weibull density function. Appendix 1 gives further details.

The left hand side variable, the log of the survival times after age 65 ($t(i)$), uses the full data available up to age 74 -- the last age before all observations are right censored (measured to twelfths of a year; recall Figure 2).

The main covariate is simply average earnings from age 52 (the earliest year common to all the observations; recall Figure 2) to the last year before retirement (i.e. up to but not including the last year of non-zero earnings).[8] Note that this represents a change from the earlier graphical results where average earnings included all available years of earnings (for some observations extending back to age 43). Using earnings only from age 52 on puts all the observations on a common footing. For example, the length of the post-65 follow-up period will not have duration until right censoring correlated with inclusion of earnings at ages below age 52, thereby avoiding a possible confounding effect. Also, since separate regressions are estimated for each age at retirement, the period over which earnings are averaged is the same for each regression.

Two other covariates have been included in the regressions essentially to provide adjustments. One is the percentage of the years included in computing average (updated) earnings where earnings were below $2,500. These percentages are intended to capture the non-monotonicity in the mortality gradient at very low incomes as is evident in Figure 1, for example.

The other covariate is the percentage of years where earnings appear to have been top-coded (i.e. truncated) at $9,999 (current dollars). This occurred from 1966 to 1971 and affects the age 52+ earnings histories of 46,936 persons and 88,111 person-years of earnings out of the total of 5,547,042 person-years included in the regressions -- i.e. about 1.6% of the person-years of earnings. In the most recent year of top-coding, 1971, when the effect of the nominal $9,999 upper limit would have been most pervasive, it appears that about 25% of contributors were affected.

To fit the model, the procedure LIFEREG from SAS (1985, pp. 507-528) was used. This program uses Newton-Raphson techniques to find maximum likelihood estimates of the coefficients.

---

[8] A time trend has not been included because it would be colinear with average earnings. There has been a significant decrease in the number of males retiring at age 65 for the population being studied, but this does not have a confounding effect because of the use of separate regressions for each age at retirement.

Because of their significantly different survival patterns, the model was fitted only to individuals who had never received a disability benefit.

An advantage of a proportional hazards model over a simple logistic model (e.g. as used by Duleep (1989), Wigle et al. (1989), Marmot (1986)) is that it uses more than just the binary information as to whether or not the individual died during the period of observation. The model uses both the length of time he lived after age 65, if he died before the last date of observation (September 1988), or the fact that he survived past this date. The CPP data give up to 100 months of survival information for each individual, so it is clearly desirable to use this information.

The validity of the Weibull regression model was checked graphically as described in Appendix 1 (equation 11) and by comparison of the residuals to the extreme value distribution. As well, the specification was checked for the assumption of linearity with respect to average earnings by further analysis of residuals.

Figure 7 shows the results of these 26 Weibull regressions, focussing on the effects of marital status and age at retirement. The histogram at the bottom shows the proportion of the population who retired at each single year of age by marital status. (Recall that retirement is here defined as the year after the last year of non-zero earnings.) About half the population retired in the year they attained age 65.

The two solid lines show the average survival probabilities from age 65 to age 70 by year of retirement and marital status. These probabilities are computed from the estimated regressions assuming average earnings are \$25,000.[9] The dashed lines give 95% "confidence intervals"[10].

Consistent with earlier results, married males have significantly higher survival probabilities at all retirement ages. More interestingly, and in line with Figure 6, there appears to be a somewhat monotone increasing relationship between survival probability and age at retirement. However, the patterns are not entirely uniform or parallel for the two marital states. Thus, separate regressions appear to have been warranted.

Figure 8 is identical to Figure 7 except that instead of showing confidence intervals, a second pair of curves is shown based on earnings of \$50,000. The difference between the two solid lines thus shows the effects on survival probabilities of an increase in average pre-retirement earnings from \$25,000 to \$50,000 for married men by age at retirement. The difference between the two dashed lines shows the corresponding effects for non-married men. Higher earnings always entail higher survival probabilities; but the magnitude of this earnings gradient tends to narrow for later retirement ages. The effect is similar but somewhat more variable among not married men.

Figure 9 illustrates the implications of the regressions in terms of relative risks. Instead of showing only two earnings levels as in Figure 8, this graph shows the impacts for married men at various percentiles of the overall average earnings distribution (i.e. for the distribution of average earnings pooled for all retirement ages and both marital statuses). For the most numerous group, married men retiring at age 65 (208,572 observations -- see

---

[9] The probabilities also assume that there were no years with top-coded earnings, and no years with very low earnings. Appendix 2 gives details of the regressions. As shown in Appendix 2, there is no systematic bias in the results associated with the variables for top-coding, or with very low earnings.

[10] More precisely, these are the transformed confidence intervals for the log of life length.
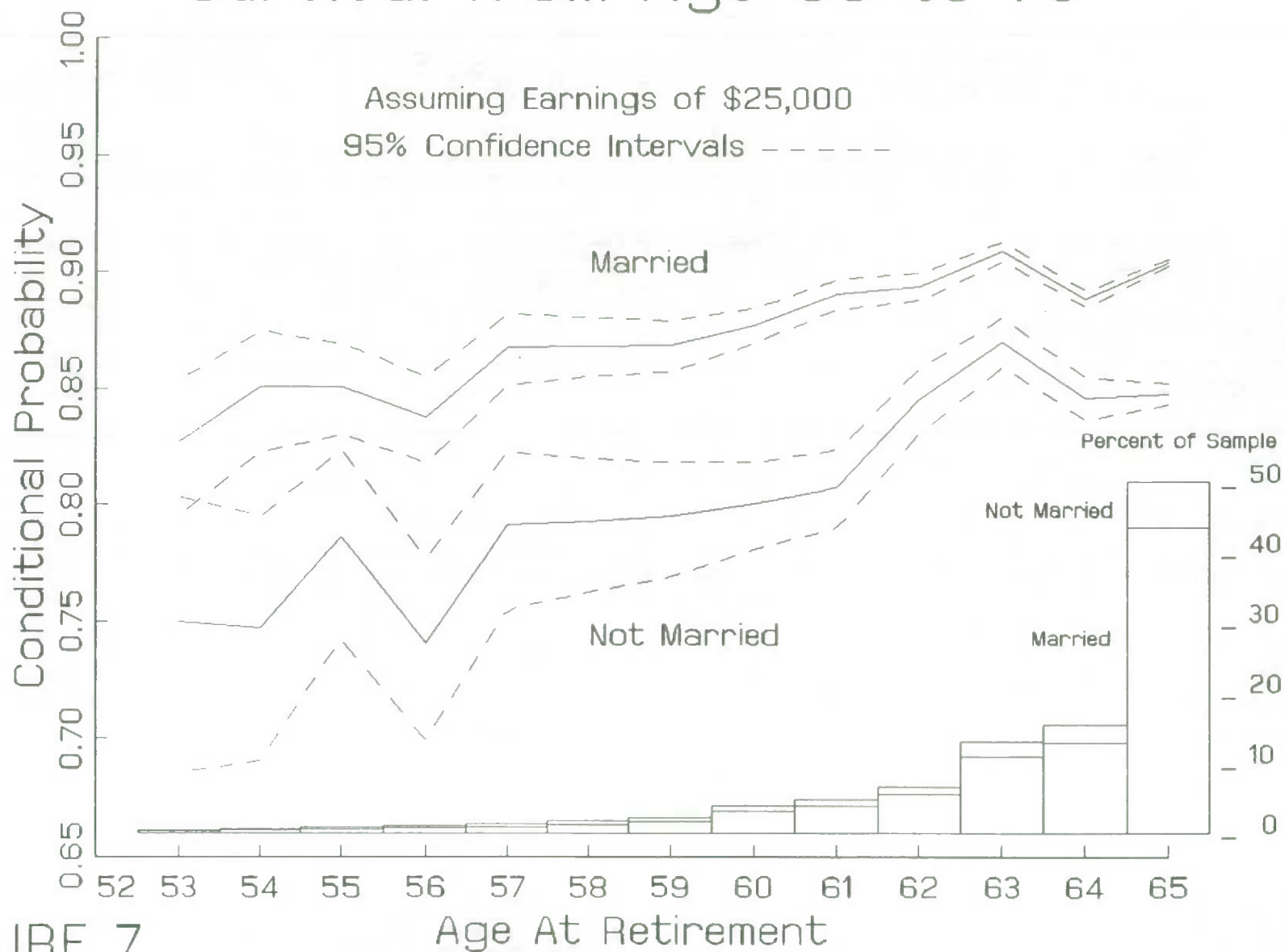
Survival from Age 65 to 70

FIGURE 7

Survival from Age 65 to 70

By Earnings, Marital Status & Age at Retirement
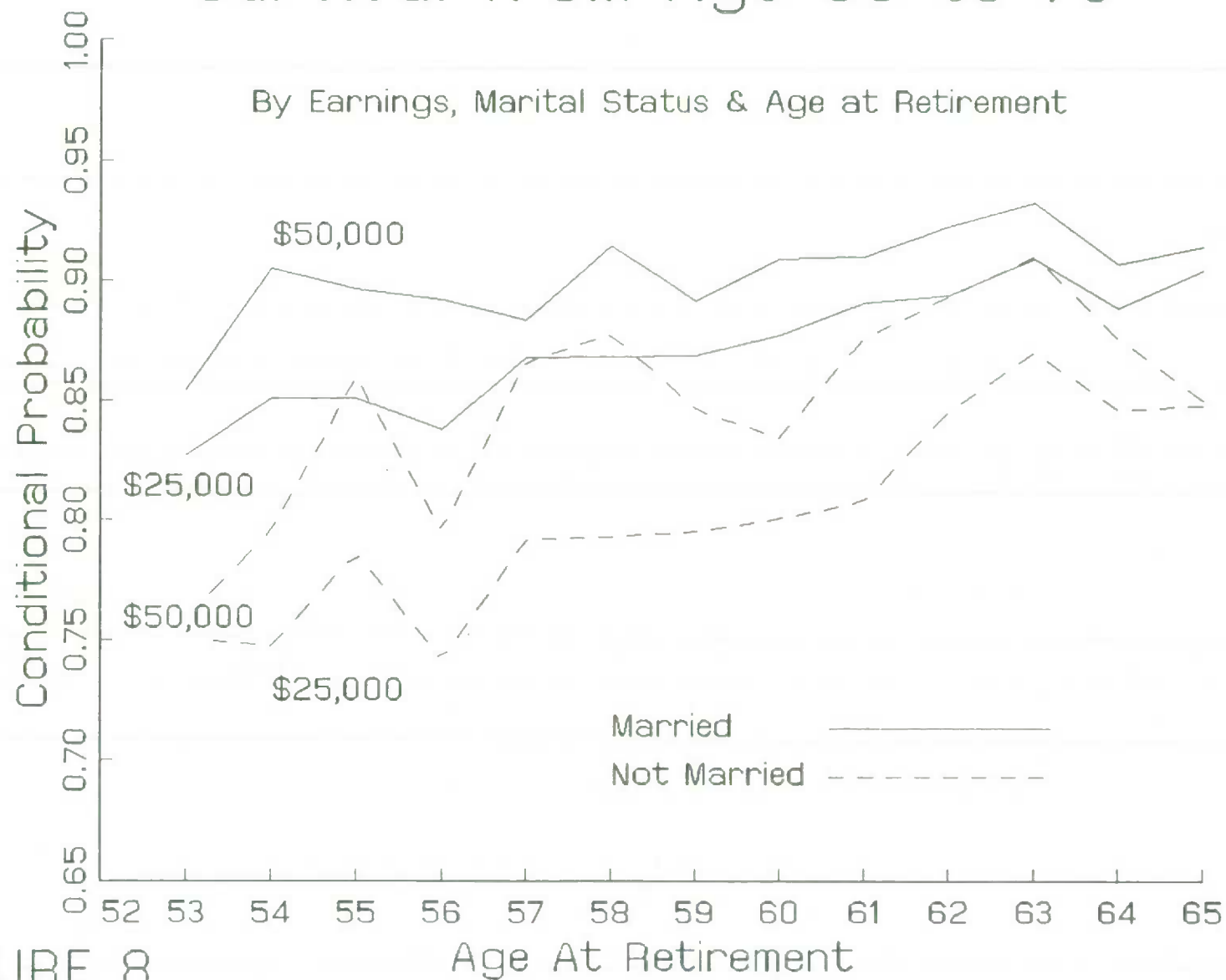
FIGURE 8

Relative Risk at Age 70
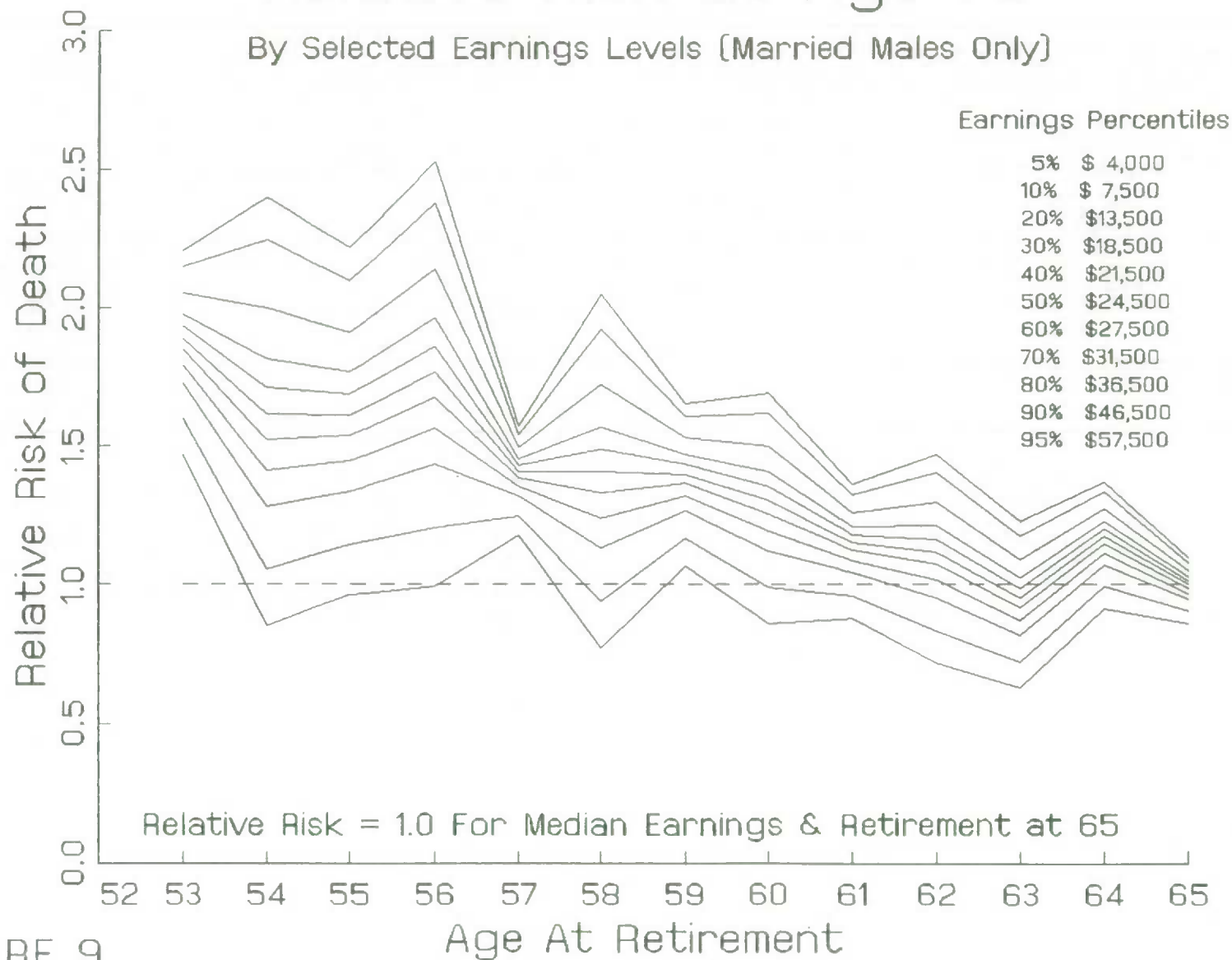By Selected Earnings Levels (Married Males Only)

Earnings Percentiles

5%	$ 4,000
10%	$ 7,500
20%	$13,500
30%	$18,500
40%	$21,500
50%	$24,500
60%	$27,500
70%	$31,500
80%	$36,500
90%	$46,500
95%	$57,500

Relative Risk = 1.0 For Median Earnings & Retirement at 65

FIGURE 9

Appendix 2), relative risks range from .86 for the 95th percentile to 1.10 for the 5th percentile, where the relative risk for median average earnings (equal to about $24,500) has been set to 1.00.

With earlier retirement, relative risks at median earnings rise to over 1.5. At the same time, the range of relative risks across earnings percentiles widens to over twice as high for the 5th compared to the 95th earnings percentile.

H. Discussion

A number of significant results have been derived from this multivariate regression analysis. One concerns the shape of the earnings gradient. Since the Weibull regression is based on a linear relationship between log life length and average earnings, the survival probability over a given age interval is of the form given in equation (11) in Appendix 1. Thus, an extra dollar of income offers decreasing "protective effect" at higher income than at lower incomes; an intuitively plausible result.

As well, since average earnings is itself a function of earnings in all the years between age 52 and the year before retirement (inclusive), an extra dollar of earnings in any of these years has the same "protective effect."[11] This may be more surprising intuitively, though it accords with the notion that "permanent" rather than "transitory" earnings is the key variable. In turn, this suggests that there are long term effects of earnings on mortality, with lagged associations of as much as decades. It also suggests that not only cross-sectional analyses but also shorter term (e.g. 2 years) mortality follow-up studies using an annual income variable such as Rogot et al. (1988) may miss or understate important relationships.

The magnitude of the simple univariate earnings gradient shown in Figure 1 is reduced and becomes variable when account is taken of other factors in a multivariate analysis. When those claiming CPP disability benefits are excluded, the variations in mortality with respect to marital status and age at retirement are of the same general order of magnitude as the gradients with earnings in various sub-groups. Still, all these variations are non-trivial. Expressed in terms of relative risks, the impacts on post age 65 mortality of variations in pre-retirement average earnings, marital status, and age at retirement are of the same order as the impact of smoking or high blood cholesterol levels on the risk of a heart attack (i.e. relative risks of 1.5 to 2.0; e.g., Wilson et al. (1987), Semenciw et al. (1988)).

The increase in mortality associated with earlier ages at retirement suggests some sort of health effect. Onset of illness may predispose an individual both to withdraw from work -- retire earlier than age 65, and to higher mortality after age 65. The notion of a "health selection" effect has been used to argue that gradients in mortality with respect to social class are artifactual, the result of poor health causing both lowered earnings (or social class) and higher mortality.

---

[11] Other regressions not reported here used earnings in each year from age 52 to the year before retirement as right hand side variables in order to explore the effects of the "shape" of the pre-retirement earnings profile. Except for age at retirement, and trend in earnings as mentioned below, these more detailed regressions did not add importantly to explanatory power. This is in accord with the negligible differences between the "interrupted" and "uninterrupted work history" earnings gradients shown in Figure 6.

While a "health selection" effect may be operative for some of the population studied here, it is clearly not applicable to everyone. This was shown in Figure 9. Mortality gradients in relation to earnings are evident within groups who retired at the same age; indeed they are larger for earlier ages at retirement.

As well, it is implausible for this kind of "health selection" effect to be operating among those whose earnings were generally increasing (relative to the average wage, not just in nominal or real terms) prior to retirement. Yet, as shown in Figure 4, mortality gradients with average pre-retirement earnings are still apparent for this sub-group.

Further evidence on this point is provided by an additional regression, reported in Appendix 2. This regression is the same as that already displayed in Figures 7 to 9 for married males retiring at age 65 (over 200,000 observations -- the modal group comprising almost 40% of the study population) with one exception. The regression included one additional variable -- the rank correlation of pre-retirement earnings and age (the same variable that was used to select the sub-population in Figure 4).

The coefficient for this term indicates a statistically significant and non-trivially positive association with post-retirement survival duration. Thus, holding average pre-retirement earnings, age at retirement, and marital status fixed, and excluding the "disabled", an increasing trend in pre-retirement earnings was associated with enhanced survival probabilities after age 65 (and vice versa). Intuitively, this seems to suggest that when "things are getting better" economically, this has a beneficial effect on survival many years later, regardless of any health effects just prior to retirement.

Thus, to recapitulate our findings with regard to the "health selection" hypothesis, we have

- excluded the seriously disabled;

- used average earnings, thereby minimizing the impacts of any acute health conditions;

- excluded earnings in the year of retirement; thereby excluding years likely to have been affected by any critical health events;

- disaggregated by age at retirement;

- controlled for the effects of chronic degenerative health effects to the extent they limit earnings by including in the analysis individual level trends in earnings relative to average wages; and

- considered associations between earnings and mortality where the lags are quite long -- earnings between ages 52 and the early 60s, and mortality between age 65 and 74 conditional on surviving to age 65.

Even with all these considerations, a significant gradient in mortality as a function of earnings is apparent. Thus, it is highly unlikely that health selection is the sole explanation.

As noted in the beginning sections of the paper, the existence of mortality gradients with marital status and earnings (or related socio-economic status variables) is generally well known, though typically without the detail and tight confidence intervals presented here. Explanations, however, are much less certain. One important explanation -- health selection -- is evidently not plausible for large segments of the population studied here.

Other results in this analysis such as the positive association of post age 65 survival duration with increasing trends in pre-retirement earnings, with retirement age, and the widening gradient with earnings at earlier retirement ages, holding other things fixed, appear to be new. The first of these results has an intuitive appeal as just noted. There are some suggestions as to the causal pathways, for example as cited earlier from Hirdes et al. (1986); but they are certainly not clear.

Intuition or explanations in the latter two cases are more problematic. The literature on retirement behaviour suggests industry and occupation, level of earnings, health status, and expected levels of pension benefits as key determinants. Following Burtless (1987) for example, one possible explanation for the association with retirement age is that retirement is customarily earlier in occupations or industries that are most demanding in terms of adverse health effects (e.g. the "30 years and out" normal retirement in some industrial pension plans).

However, the extent of early retirement in the CPP data (recall the histogram in Figure 7) appears quite low -- for example compared to the "normal" retirement ages in private pension plans, as well as the subsidized (in actuarial terms) "special early retirement" provisions and anecdotal evidence since the early 1980s of the use of "golden handshakes" and early retirment to assist in "downsizing" (Statistics Canada, 1989).

The results of the analysis thus raise important questions about causal mechanisms. Conventional explanations cannot easily account for some of the results.

In addition, the results also have important implications for public policy. In health policy, the existence of significant gradients raises questions about the efficacy of the current health insurance system. There are two broad possibilities. On the one hand, the health care system might not be offering equal access given need. Males with lower average earnings histories may be receiving poorer quality care, even though they visit health care providers with the same frequencies in relation to the prevalence of health problems (Manga et al., 1987; Broyles et al, 1983).

Alternatively, there may be aspects of lifestyle, work place, or home that vary systematically with earnings, that also predispose to higher mortality, and that are not affected by the services offered by the health care system. For example, low income workers may be exposed to higher levels of stress or toxins such that once they become ill, it is too late for the health care system to provide much in the way of cure.

Either hypothesis raises serious though quite different concerns. Which one is most appropriate requires further research.

A second major area where these results are important to public policy is pensions. The results suggest that Canada's public pension system is not as progressive as many think. In lifetime income terms, if higher income individuals live longer, they collect pensions for a longer period. Thus, the earnings-related CPP which appears distributionally neutral because of its constant 25% replacement rate is actually regressive. (This point has been made for Blacks and Whites in the context of the U.S. Social Security system by Schulz (1974).)

Similarly, the gradients in post age 65 mortality by age at retirement raise questions about the equity of current actuarial adjustment factors for the recently legislated early retirement benefits under the C/QPP. These actuarial factors are roughly "neutral" under

the assumption that mortality rates do not depend on age at retirement. Thus, early retirees who, based on the results presented above, appear to face higher mortality prospects will receive smaller lifetime C/QPP benefits.

I. Summary and Conclusions

This study has examined the relationship between pre-retirement earnings histories and mortality after age 65 for over half a million Canadian males. The data show a clear and significant gradient -- higher earnings decades prior to age 65 are associated with lower mortality during the following nine years. As well, being married, not retiring early, not being disabled, and having improvements in earnings are all significantly associated with higher survival probabilities.

On a methodological note, these kinds of results illustrate the as yet largely unexploited power of administrative data for social science and medical research.

Finally, the causal pathways by which these socio-economic status variables may influence mortality are generally unknown. However, juxtaposing the gradient in mortality with the generally equal access to medical care services in Canada, without regard to financial position, raises fundamental questions about the most important directions for health research.

References

Blaxter, M. (1986), "Longitudinal studies in Britain relevant to inequalities in health", in R.G. Wilkinson (Ed.), *Class and Health, Research and Longitudinal Data*, Tavistock Publications, London.

Broyles, R.W., P. Manga, D.A. Binder, D.E. Angus and A. Charette (1983), "The Use of Physician Services Under a National Health Insurance Scheme", *Medical Care*, Vol. 21, No. 11, November.

Burtless, G. (1987), "Occupational Effects on the Health and Work Capacity of Older Men", in G. Burtless (Ed.), *Work, Health, and Income Among the Elderly*, The Brookings Institution, Washington.

D'Arcy C. (1989)"Reducing inequality in health: A Canadian perspective, Part II: A report of a review of literature and data", University of Saskatchewan.

Duleep, H.O. (1986a), "Measuring the Effect of Income on Adult Mortality Using Longitudinal Administrative Record Data", *Journal of Human Resources*, Vol. 21, No. 2, Spring.

Duleep, H.O. (1986b), "Incorporating Longitudinal Aspects into Mortality Research Using Social Security Administrative Record Data", *Journal of Economic and Social Measurement*, Vol. 14, No. 2, July.

Duleep, H.O. (1989), "Measuring Socioeconomic Mortality Differentials Over Time", *Demography*, Vol. 26, No. 2, May.

Fox, A.J., P.O. Goldblatt, and D.R. Jones (1985), "Social class mortality differentials: artefact, selection or life circumstances?", *Journal of Epidemiology and Community Health*, Volj. 39, pp. 1-8.

Goldblatt, P.O. (1989), "Mortality by Social Class, 1971-85", *Population Trends*, No. 56, Summer.

Hirdes, J.P. and W.F. Forbes (1989), "Estimates of the Relative Risk of Mortality Based on the Ontario Longitudinal Study of Aging", *Canadian Journal of Aging*, Vol. 8, No. 3.

Hirdes, J.P., K.S. Brown, W.F. Forbes, D.S.Vigoda, and L. Crawford (1986), "The Association Between Self-Reported Income and Perceived Health Based on the Ontario Longitudinal Study of Aging", *Canadian Journal of Aging*, Vol. 5, No. 3.

Isnard, M. (1989), "Mortalité et changements de catégorie socioprofessionnelle: Premiers Résultats de l'Échantilleon Démographique Permanent", communication à la 5ième réunion du réseau ONU/OMS/CICRED sur les différences socio-économiques de mortalité dans les sociétés industrialisées, Paris.

Kalbfleisch, J.D. and R.L. Prentice (1980), *The Statistical Analysis of Failure Time Data*, Wiley.

Kennedy, B. (1987), ""True" Age Profiles of Earnings: Confirmation and Extension of Hanoch and Honig", Institute for Research on Public Policy Working Paper, February.

Kitagawa, E.M. and P.M. Hauser (1973), *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*, Harvard University Press, Cambridge.

Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data*, Wiley.

Lundberg, O. (1986), "Class and Health: Comparing Britain and Sweden", Reprint Series No. 166, Swedish Institute for Social Research, reprinted from *Social Science and Medicine*, Vol. 23, No. 5.

Manga, P., R.W. Broyles and D.E. Angus (1987), "The Determinants of Hospital Utilization Under a Universal Public Insurance Program in Canada", *Medical Care*, Vol. 25, No. 7, July.

Marmot, M.G. (1986), "Social inequalities in mortality: the social environment", in R.G. Wilkinson (Ed.), *Class and Health, Research and Longitudinal Data*, Tavistock Publications, London.

McKeown, T. (1984), "Research Strategy: the Role of WHO", World Health Organization, Geneva.

Nagnur, D.N. and M. Nagrodski, (1987), "Cause-Deleted Life Tables for Canada (1921 to 1981): An Approach Towards Analysing Epidemiologic Transition", Analytical Studies Branch Research Paper Series No. 13, Statistics Canada, Ottawa.

Rogot, E., P.D. Sorlie, N.J. Johnson, C.S. Glover and D.W. Treasure (1988), "A Mortality Study of One Million Persons by Demographic, Social, and Economic Factors: 1979-1981 Follow-up", U.S. Department of Health and Human Services, NIH Publication No. 88-2896, March.

SAS Institute Inc. (1985), *SAS User's Guide, Statistics*, Version 5 Edition.

Schulz, J. et al. (1974) *Providing Adequate Retirement Income: Pension Reform in the U.S. and Abroad*, Brandeis University Press, Hanover N.H.

Semenciw, R.M., H.I. Morrison, Y. Mao, H. Johansen, J.W. Davies and D.T. Wigle (1988), "Major Risk Factors for Cardiovascular Disease Mortality in Adults: Results from the Nutrition Canada Survey Cohort", *International Journal of Epidemiology*, Vol. 17, No. 2.

Statistics Canada (1989), *Pension Plans in Canada*, Catalogue 74-401, Ottawa.

Statistics Canada (1980), *Vital Statistics, Volume III -- Deaths, 1977*, Catalogue 84-206, Ottawa.

Townsend, P. and N. Davidson (eds.) (1988), *The Black Report*, The Penguin Group, London.

Wigle, D.T. and Y. Mao (1980), *Mortality by Income Level in Urban Canada*, Health and Welfare Canada, Ottawa.

Wigle, D.T., Y. Mao and G. Arraiz (1989), "Mortality Follow-Up Study: Results from the Canada Health Survey", paper presented to the Canadian Epidemiology Conference, University of Ottawa, Ottawa, August.

Wilkins. R., O. Adams and A. Brancker (1989), "Mortality by Income in Urban Canada, 1971 and 1986: Diminishing Absolute Differences, Persistence of Relative Inequality", Joint Study, Health and Welfare Canada and Statistics Canada, Ottawa.

Wilson P.W.F., W.P. Castelli, and W.B. Kannel (1987), "Coronary Risk Prediction in Adults (The Framingham Heart Study), *Americal Journal of Cardiology*, Vol. 59, pp. 91G-94G.

Appendix 1: Description of the Weibull Regression Model

The Linear Regression Model and its Relation to the Weibull Distribution

Let the random variable T be lifelength, and let $Y = \ln[T]$. The observed data for n individuals are $t_1, t_2, ..., t_n$ (or $y_1, y_2, ..., y_n$, where $y_i = \ln[t_i]$). Assume, for the time being, that all of the t_i's are uncensored. A linear model is assumed for the y_i's:

(1) $y_i = x\,b + s\,e_i,$

where x is a row vector of covariates, b is a column vector of unknown coefficients, s is an unknown scale parameter, and e_i has the standard extreme value distribution.

It will be shown that Eqn. (1) implies that the t_i's have a Weibull distribution. The Weibull distribution is frequently used in analysis of human lifetime data because of certain properties which make it realistic for representing such data.

The extreme value probability density function (PDF) for the e_i's is

(2) $f_e(e) = \exp[\,e - \exp[e]\,]$ ($-\infty < e < \infty$).

Therefore, the PDF for the y_i's is[12]

(3) $f_y(y) = (1/s)\,\exp[\,(y - x\,b)/s - \exp[\,(y - x\,b)/s\,]\,]$ ($-\infty < y < \infty$).

(This is obtained using the usual change of variable relationship that $f_y(y) = f_e(e)\,de/dy$ and substituting $(y - x\,b)/s$ for e and $1/s$ for de/dy.) From Eqn. (3), the PDF $f_t(t)$ for the t_i's is obtained:

(4) $f_t(t) = (1/(t\,s))\,\exp[\,(\ln[t] - x\,b)/s - \exp[\,(\ln[t] - x\,b)/s\,]\,].$

(This is obtained using the change of variable relationship that $f_t(t) = f_y(y)\,dy/dt$ and substituting $\ln[t]$ for y and $1/t$ for dy/dt.) Eqn. (4) is the PDF of a Weibull distribution. It can be rewritten more familiarly as follows:[13]

(5) $f_t(t) = (1/(s\,\exp[\,x\,b\,]))\,(t/(\exp[\,x\,b\,]))^{(1-s)/s}\,\exp[\,-t/(\exp[\,x\,b\,])^{1/s}\,]$ ($t > 0$)

The parameters b and s are estimated by maximizing the likelihood $TT_i\,f_y(y_i)$ of the y_i's or, equivalently, by maximizing the likelihood $TT_i\,f_t(t_i)$ of the t_i's.

Weibull Regression as a Proportional Hazards Model

The distribution of a random variable is uniquely and completely identified by its PDF $f(.)$ or by its cumulative distribution function (CDF) $F(.)$. In analyzing lifetime data, the survival distribution $S(.) = 1 - F(.)$ and the hazard function $h(.) = f(.)/S(.)$ are also used to identify a distribution. For the Weibull distribution having PDF $f_t(t)$ as in Eqn. (5), the survival distribution $S(t)$ is

(6) $S(t) = \exp[\,-(t/\exp[\,x\,b\,])^{1/s}\,].$

---

[12] cf. Lawless (1982, p. 274, Eqn. 6.1.3).

[13] cf. Lawless (1982, p. 141), with alpha = $\exp[\,x\,b\,]$ and beta = $1/s$.

Thus, the hazard function of t (given x) is[14]

(7) $h(t;x) = (1/(s \exp[x b])) (t/\exp[x b])^{(1-s)/s}$.

Note that Eqn. (7) implies that given x b, there is a simple linear relationship between the log of the hazard function and the log of the lifetime:

(8) $\ln[h(t;x)] = -\ln[s] - x b/s + ((1-s)/s) \ln[t]$.

The ratio of the hazard functions for two individuals with covariate vectors $x=x_1$ and $x=x_2$ is

(9) $h(t;x_1) / h(t;x_2) = (\exp[(x_2 - x_1) b])^{1/s}$,

which does not depend on the lifelength t. Thus, different individuals have "proportional hazards", irrespective of t. (Proportional hazards models may also be based on other hazard functions than a Weibull hazard.)

Relative Risk

If two individuals differ in only one covariate, then the ratio of their hazard functions reduces to

(10) $h(t;x_1) / h(t;x_2) = (\exp[(x_{2k} - x_{1k}) b_k])^{1/s}$,

where x_{1k} and x_{2k} are the values of that covariate for the two individuals, and b_k is the coefficient for that covariate. Eqn. (10) is used to assess the "relative risk" associated with a change from x_{1k} to x_{2k} in the k-th covariate, holding constant the values of all other covariates.

Censored Data

If some of the t_i's are censored, the likelihood function used to estimate the regression parameters changes from $TT_i\, f(t_i)$ to

$$TT_i\, (f(t_i))^{d_i}\, (S(t_i))^{1-d_i},$$

where $d_i = 1$ if t_i is uncensored, and $d_i = 0$ if t_i is censored. Thus, information about ongoing lifetimes is incorporated into the model.

Checking the Weibull Assumption

From Eqn. (6), it is seen that

(11) $\ln[-\ln[S(t)]] = (\ln[t] - [x b]) / s$.

Thus, an empirical check for the validity of the Weibull distributional assumption is provided by a plot of $\ln[-\ln[S(t)]]$ (using an estimate of S(t)) versus $\ln(t)$. Under the Weibull assumption, this plot should yield an approximately straight line configuration of points. The slope of the line provides a rough estimate of 1/s, and the vertical axis intercept can be used to estimate $-\exp[x b]/s$ (and thus x b). (See Kalbfleisch and Prentice, 1980, p. 24.)

---

[14] cf. Lawless (1982, p. 274, Eqn. 6.1.2, with delta = 1/s and alpha(x) = exp [x b].

The quantity -ln[S(t)] is sometimes called the "cumulative hazard" because it is equal to the integral of the hazard function from zero to t.

Appendix 2: Coefficients of the Weibull Regressions

The following table contains the estimated coefficients for 27 Weibull regressions (described in Appendix 1), with standard errors shown beneath the coefficients.

The first half of the table is for the not married male population; the second half is for married males.

Within each marital status group, there are 13 sets of regression results, one set for each age at retirement. In one case, married with retirement age 65, there is a second regression that contains one extra right hand side variable.

The left hand side variable in all cases is the log of life length after age 65 in years, measured to the nearest twelfth. The right hand side variables are as follows:

Const -- constant term

Earn -- average earnings in tens of thousands of dollars, where the average has been computed for each individual by first "updating" each year's earnings by multiplying it by the ratio of the average wage in 1988 to the average wage in the year of the earnings, and then averaging these "updated" figures; only earnings between age 52 and the year before the last year of non-zero earnings are included

Top -- fraction of all earnings years prior to retirement that were top-coded

Low -- fraction of all earnings years prior to retirement that were below $2,500

Tau -- rank correlation between age and earnings from age 52 to 64 (used in only one regression)

Also shown below are

Scale -- scale parameter for the extreme value distribution of errors

Deaths -- the number of persons in the regression where the individual died before the end of the period of observation, September 1988

Censored -- the number of individuals in the regression who were still alive at the end of the period of observation

Note that the number of observations in each regression is the sum of Deaths and Censored. All coefficients are significant at the 5% level unless other wise noted.

NOT MARRIED MALES

Age	Const	Earn	Top	Low	Scale	Deaths	Censored
65	2.90440	0.00489**	0.10150**	0.33234	0.72586	5,061	26,718
	0.02568	0.00810	0.22451	0.07342	0.00897		
64	2.99524	0.08278	-1.31228	0.05734**	0.89087	1,512	10,568
	0.06166	0.02295	0.52478	0.13360	0.01952		
63	2.85259	0.12157	-1.13415	0.27225*	0.78360	789	8,925
	0.07686	0.02826	0.57473	0.16371	0.02243		
62	2.72166	0.12779	-0.96846**	-0.11642**	0.80232	673	4,698
	0.08833	0.03451	0.59088	0.16820	0.02567		
61	2.45090	0.15154	-0.03489**	0.21041**	0.79048	681	3,627
	0.08269	0.03346	0.59000	0.16395	0.02518		
60	2.69568	0.06887*	-0.59665**	0.07680**	0.83735	605	3,190
	0.09372	0.03587	0.47917	0.17737	0.02853		
59	2.55392	0.10399	-0.37460**	0.10940**	0.81768	474	2,230
	0.09883	0.04337	0.59787	0.18324	0.03120		
58	2.39547	0.20364	-0.49880**	0.35796*	0.88631	410	1,849
	0.10929	0.05191	0.56387	0.19590	0.03694		
57	2.39587	0.15963	-0.08590**	0.13195**	0.81564	345	1,539
	0.11027	0.05485	0.77344	0.18122	0.03661		
56	2.43017	0.09488*	0.65209**	0.52989	0.87821	299	1,286
	0.11582	0.05512	0.65827	0.21761	0.04256		
55	2.46218	0.16539	-0.58390**	0.38164*	0.88833	234	1,101
	0.13254	0.06617	0.54961	0.22060	0.04877		
54	2.50952	0.08724**	0.31101**	-0.02196**	0.90630	204	810
	0.13722	0.07163	0.60058	0.16263	0.05246		
53	2.77162	0.01582**	0.15024**	-0.14366**	0.96398	134	688
	0.17571	0.07615	0.56726	0.20684	0.07000		

MARRIED MALES -- Special Regression Including Age-Earnings Correlation

Age	Const	Earn	Top	Low	Scale	Tau
65	3.32337	0.03154	-0.00037**	0.19229	0.77991	0.07055
	0.01518	0.0039	0.08995	0.04297	0.00483	0.01362

MARRIED MALES

Age	Const	Earn	Top	Low	Scale	Deaths	Censored
65	3.31247	0.03527	-0.05910**	0.22369	0.78027	20,952	187,620
	0.01503	0.00385	0.08944	0.04259	0.00483		
64	3.48937	0.07334	-0.07547**	0.14335**	0.96559	4,974	56,825
	0.03891	0.01118	0.24986	0.09151	0.01176		
63	3.24130	0.09997	-0.52842	0.10265**	0.80071	2,339	49,890
	0.04900	0.01408	0.25048	0.10743	0.01299		
62	3.13113	0.10976	-0.69356	0.00055**	0.82092	1,736	24,170
	0.06005	0.01797	0.25159	0.12206	0.01628		
61	3.17650	0.06597	-0.10847**	-0.22602*	0.80305	1,462	16,754
	0.06297	0.01763	0.23771	0.12547	0.01771		
60	3.10184	0.10981	-0.31832**	0.23310**	0.86905	1,283	13,634
	0.07037	0.02078	0.24659	0.15512	0.02068		
59	3.15715	0.07214	0.06543**	-0.03179**	0.88101	732	7,008
	0.08941	0.02534	0.31790	0.16942	0.02773		
58	2.91643	0.15924	-0.23951**	0.16829**	0.87003	457	5,124
	0.10057	0.03308	0.31344	0.17913	0.03142		
57	3.19869	0.04761**	0.35924**	-0.29802**	0.87462	390	3,700
	0.11784	0.03197	0.33986	0.19735	0.03747		
56	2.71680	0.14939	-0.12294**	0.44377	0.85511	368	2,879
	0.11118	0.03890	0.27263	0.21731	0.03760		
55	2.77631	0.12688	0.09831**	0.33820**	0.81377	2276	2,258
	0.11726	0.04220	0.27500	0.21984	0.04097		
54	2.83806	0.17719	-0.02045**	-0.01221**	0.91712	197	1,348
	0.14704	0.05989	0.33849	0.18593	0.05524		
53	2.99709	0.07161**	0.58933**	0.19831**	0.94371	1149	1,104
	0.16905	0.05037	0.37335	0.24169	0.06676		

*   0.1 ≥ p > 0.05 (i.e. questionable significance)
** p > 0.1 (i.e. insignificant)

ANALYTICAL STUDIES BRANCH
RESEARCH PAPER SERIES

For further information, contact the Chairperson, Publications Review Committee, Analytical Studies Branch, R.H. Coats Bldg., 24th Floor, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, (613) 951-8213.