Extrapolation of the U.S. Life Tables

Terry M. Therneau, Ph.D.
Christopher Scheib

Technical Report No. 55
October 1994

I. Introduction

The routines for calculating expected survival, much used at Mayo, depend on various U.S. and state rate tables which are prepared by the Public Health Service. These tables, however, are not current. The 1980 rates were published in 1985 and the 1990 rates are not yet available.

Figure 1 shows the rates over time for 41-60 year olds from the West North Central population; it is one of many from an earlier technical report (1). The immediate impression is the steady and regular decline in the death rate over the decades. At present, both the SAS %survexp macro and the S survexp function use linear interpolation between the decade years and replication outside of the range of the tabled data. In essence, this means that the curves in Figure 1 are extrapolated to the right as horizontal lines; this is illustrated in Figure 2.

The question is not whether to extrapolate the data, as this must be done whenever a 1981 or later rate is desired, but whether a better extrapolation can be found than the one currently performed. From the linear nature of Figure 2 the answer is certainly yes. This report contains the results and recommendations of such an investigation.

II. Results

The West North Central population shown in Figure 1 is exceptional in its time span, far less data is available for the other populations under consideration:

1950, 60, 70 and 80:	U.S. and Minnesota white
1960, 70, and 80:	U.S. total and non-white
1970 and 80:	Minnesota total, Florida total, white and non-white, Arizona total and white
1980 only:	Arizona non-white, U.S. black, Florida black

Separate tables for the WNC region were last published in 1960; the 1970 and 1980 data in Figures 1 and 2 is actually the Minnesota rates. Since the 1990 WNC data set will also be filled in with Minnesota rates when they appear direct extrapolation of the WNC data set is not appropriate. Thus although the WNC has the longest duration and provides the visual rationale for this exercise it is not used further.

Figures 3a and 3b show the entire Minnesota white data set. This plot does not look nearly as "clean" as Figure 1, but careful examination shows that they are consistent. Focus first on the data from age 35 on. In each year's data

there is an increasing hazard with age, of fairly constant slope. This corresponds to the even spacing of the curves in Figure 1. In Figure 3 we also see that the 1980 data is separated from the other 3 years, which are tangled. A closer look at Figure 1 reveals that the values for 1950 to 1970 are not always well ordered, in spite of the long term downward trend, and are followed by a drop in hazard in 1980.

The data for ages 0-15 shows a similar pattern to the older ages but that ages 15-30 are more complex. Here there is a shift in the shape of the curve over time. (For the male curves it is tempting to blame the sharp increase in hazard in 1970 on the Vietnam conflict, but there is a similar shape change for females as well.)

The most general extrapolation method is to fit some function h = g (age, sex, calendar year) to each of the population, where g accounts for "all" of the features seen in the data. The crossing hazard pattern for ages 15-30 is difficult to explain, however, and we sought a simpler additive model. To this end several plots were drawn (not shown) with x = age and y = f(hazard) - f(1950 hazard) for various simple transformations f, among them the power family $f(x) = x^{\lambda}$. In an ideal transformation, one would obtain a set of equally spaced horizontal lines. The best transform appeared to be logarithms of the hazard. The final extrapolations were based on the following:

1. The data is fit to a model

$$\log(\text{hazard}) = f(\text{age}) + \beta_1 * \text{year} + \beta_2 * \text{age} * \text{year}$$

The function f is a restricted cubic spline with 6 to 8 knots. (A restricted cubic spline or natural spline is constrained to be linear at the endpoints.) The fits were done in SAS using the %rcspline macro to generate the cubic spline's predictor variables followed by a weighted regression model using proc glm. Ages 16 through 22 of the year 1970 were given a weight of 0.1 while all other points had weight 1. In all cases the age*year interaction term was small; the calendar year effect is somewhat less for the older ages but not greatly so. An age^2 term was deliberately omitted. Looking again at Figure 1, the sharp downturn in rates in 1980 is questionable. Given the long-term trend, it is not clear that this deceleration in rates will persist, and it is dangerous to use it in our extrapolations.

2. The extrapolated rates for 1990 and 2000 are not based on the fitted smooth f(age), but rather on the 1980 rates directly:

$$1990\ \log(\text{hazard}) = 1980\ \log(\text{hazard}) + 10*(\beta_1 + \beta_2 * age)$$

Rather than try to model the changing shape of the age 15-30 hazard as a function of year, we have settled for an estimate of the mean effect. An alternative, also reasonable, would be to use a smoothed version of the 1980 log(hazard) as the baseline. This would change the extrapolated rates very little, however.

3. For data sets which included information before 1970, the regression estimators for the age and year*age terms were taken and added to the actual log hazard rates for 1980 to obtain estimates for the years 1990 and 2000. For data sets which did not include years prior to 1970, thus yielding questionable coefficient estimates in their regression analyses, estimates were obtained from populations which were considered most similar to the population in question. The list of coefficient estimate substitutions follows:

Population:	Coefficients Used:
Arizona Total Female	U.S. Total Female
Arizona Total Male	U.S. Total Male
Arizona White Female	U.S. White Female
Arizona White Male	U.S. White Male
Arizona Non-White Female	U.S. Total Female
Arizona Non-White Male	U.S. Total Male
Florida Total Female	U.S. Total Female
Florida Total Male	U.S. Total Male
Florida White Female	U.S. White Female
Florida White Male	U.S. White Male
Florida Non-White Female	U.S. Total Female
Florida Non-White Male	U.S. Total Male
Minnesota Total Female	Minnesota White Female
Minnesota Total Male	Minnesota White Male
U.S. Non-White Female	U.S. Total Female
U.S. Non-White Male	U.S. Total Male

Figure 4 shows the final results for the set of populations. The hazard axis is on a base 10 log scale, and shows daily hazard rates. Furthermore, the hazard rates for children in their first year of life (age 0) are much greater than at any other time during their childhood. In order to present the data at age zero in the graphs

while keeping the rest of the plot easy to see, the rates for age zero were divided by 10 before being plotted. Thus, a point at age zero which shows up as -5 on the graph would really translate to -4.

Fitted Constants

	β_1	β_2
U.S. Total Female	-.01448	.000050
U.S. Total Male	-.00979	-.000015
U.S. White Female	-.01770	.000041
U.S. White Male	-.01271	.000061
Minnesota White Female	-.01746	.000020
Minnesota White Male	-.01615	.000092

For example, consider the data for Arizona total females. Regression coefficient estimators of -.01448 for age and .000050 for age*year are taken from the regression analysis of U.S. total females. So the final prediction equations for the years 1990 and 2000 look like this:

(1990 prediction) = (1980 rate) + (-.01448 * 10) + (.000050 * 10 * age)
(2000 prediction) = (1980 rate) + (-.01448 * 20) + (.000050 * 20 * age)

The only worrisome part of the fitting procedure was the choice of knot points for the regression spline; we found the shape of the resultant smooth to be surprisingly sensitive to the number and location of the knots. This was true even in portions of the curve that are nearly linear, such as the older ages. Fortunately the age and age*year coefficients are little affected by this. In a final confirmatory run where the smooth was allowed complete flexibility, i.e., the linear age term had 109 d.f., the fitted coefficients β_1 and β_2 differed by no more than ± 1 in their second significant digit.

III. Implementation

The extrapolated data has been added to the rate data sets within SAS and S. No changes were made to the computational routines. The old behavior can be obtained by subsetting the rate tables before use.

In S:

```
> dimnames (survexp.uswhite)[[[3]]
```

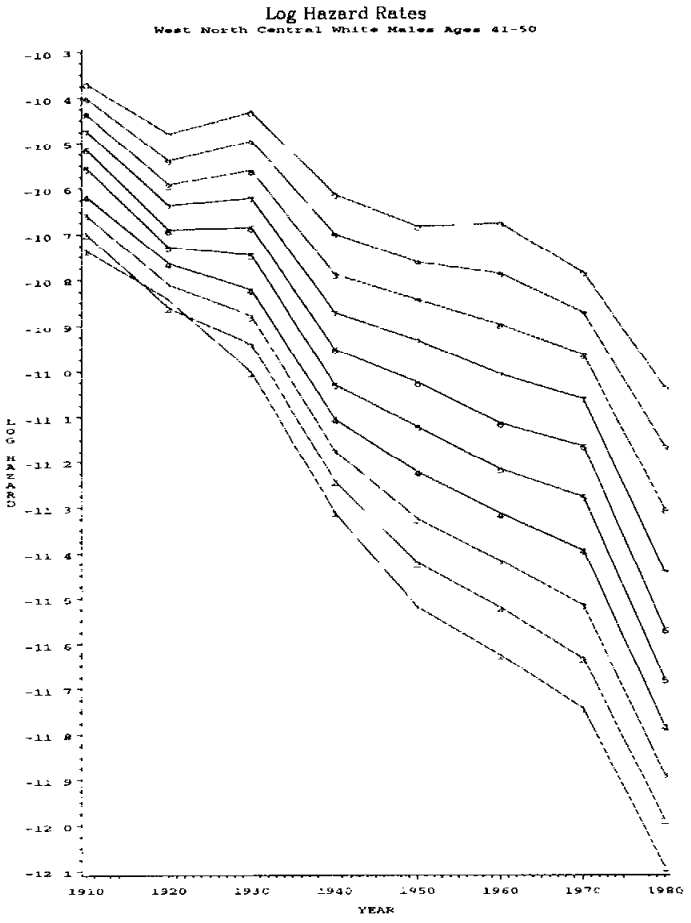
```
  1950 1960 1970 1980 1990 2000
> mytable <- survexp.uswhite [,,1:4]
> survexp (....., ratetable=mytable)
```

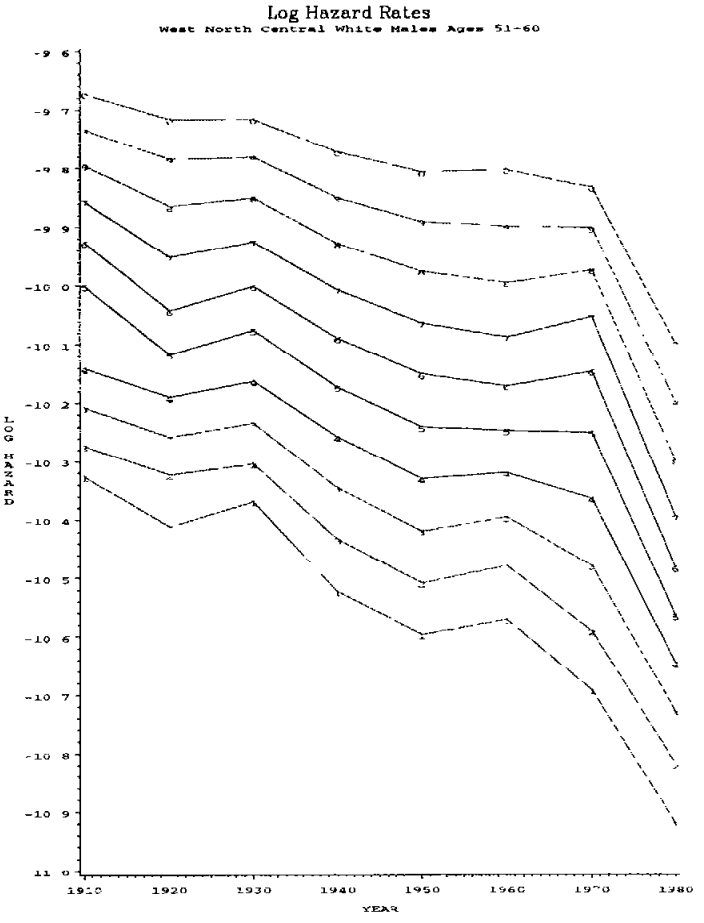
In SAS:

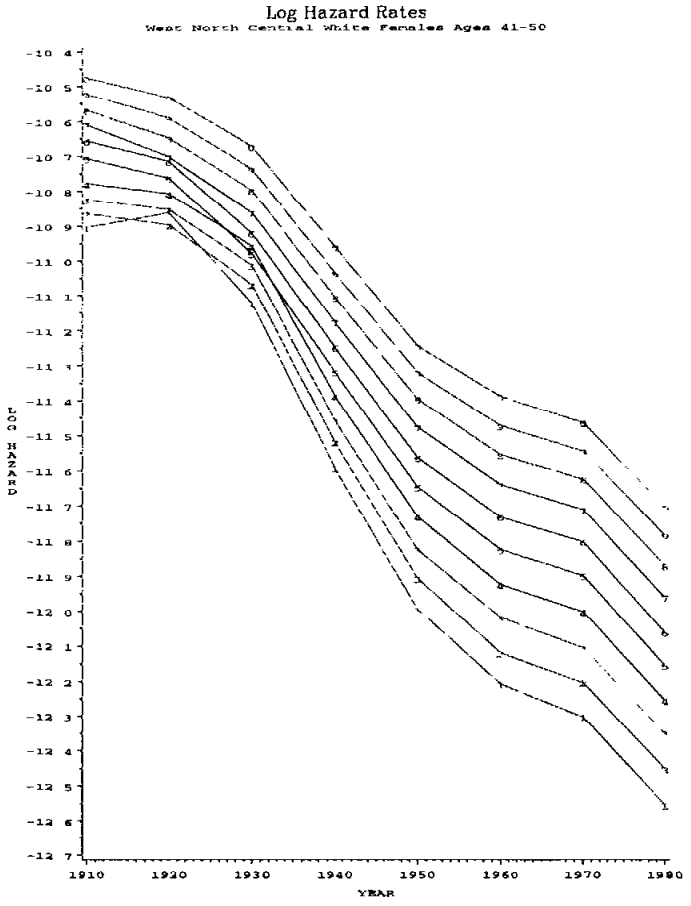
```
library rates '/usr/local/sasmac';
data lt_user; set rates.lt_us;
        if (year <= 1980);
%survexp (....., pop=user);
```

IV. References

Therneau, T.; Sicks, J.; Bergstralh, E.; and Offord, J.: Expected survival based on hazard rates. Technical Report #52, Section of Biostatistics, Mayo Clinic.


Log Hazard Rates
West North Central White Males Ages 41-50
LOG HAZARD
YEAR
1910
1920
1930
1940
1950
1960
1970
1980


Log Hazard Rates
West North Central White Males Ages 51-60
LOG HAZARD
YEAR
1910
1920
1930
1940
1950
1960
1970
1980


Log Hazard Rates
West North Central White Females Ages 41-50
LOG HAZARD
YEAR
1910
1920
1930
1940
1950
1960
1970
1980

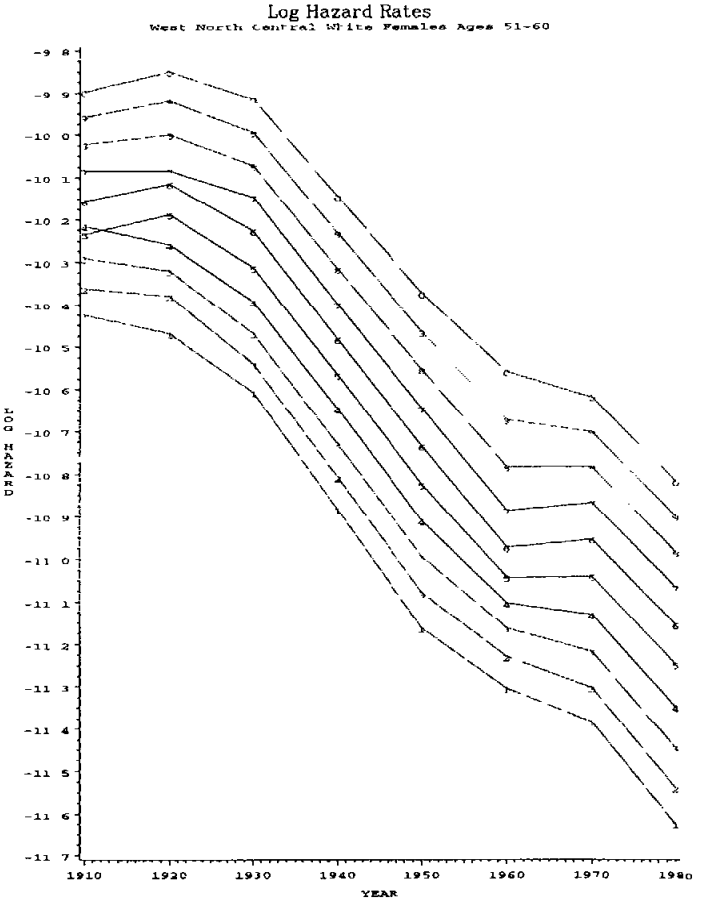

Log Hazard Rates
West North Central White Females Ages 51-60
LOG HAZARD
YEAR
1910
1920
1930
1940
1950
1960
1970
1980

Figure 1

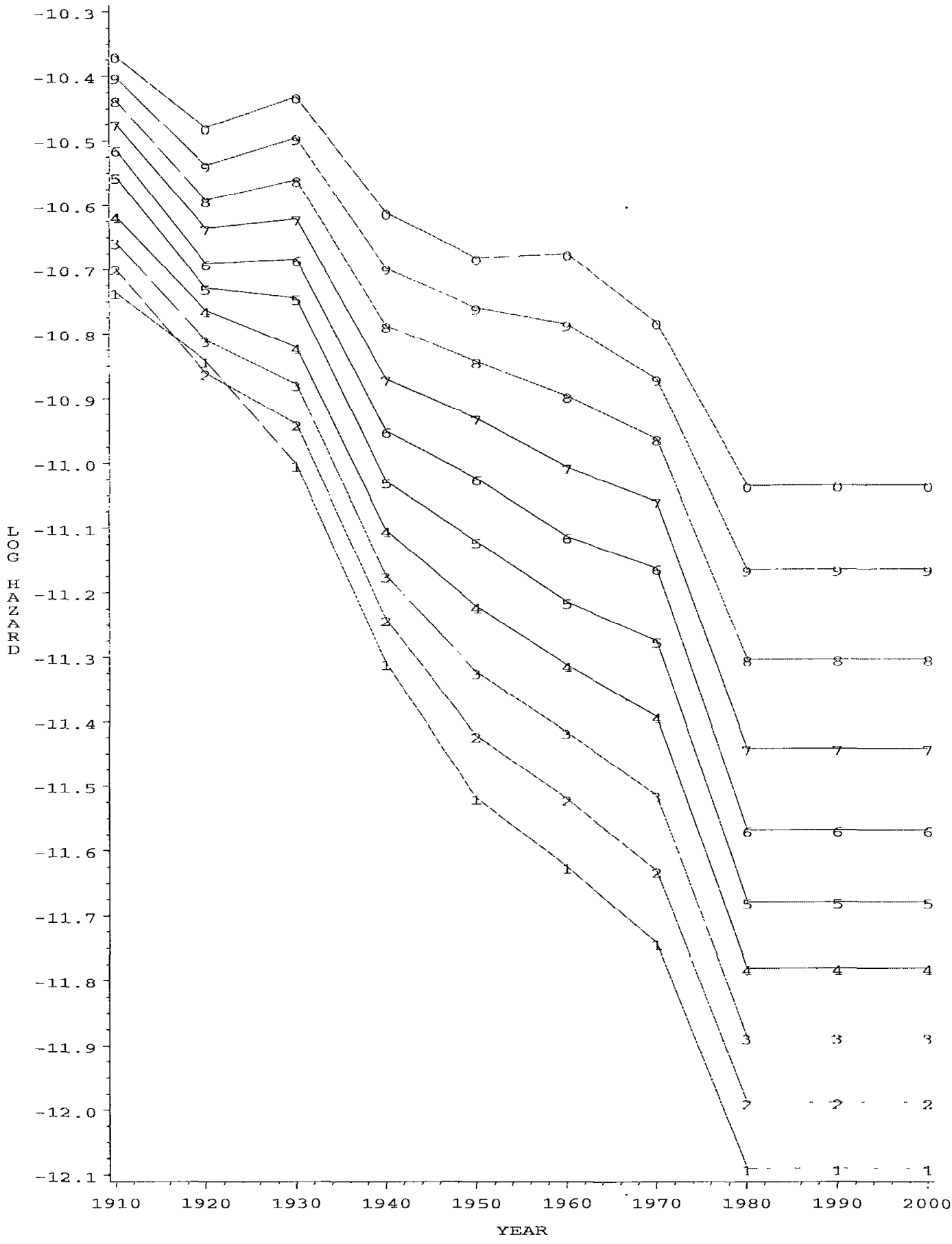

Log Hazard Rates
West North Central White Males Ages 41-50
LOG HAZARD
-10.3
-10.4
-10.5
-10.6
-10.7
-10.8
-10.9
-11.0
-11.1
-11.2
-11.3
-11.4
-11.5
-11.6
-11.7
-11.8
-11.9
-12.0
-12.1
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
YEAR

Figure 2

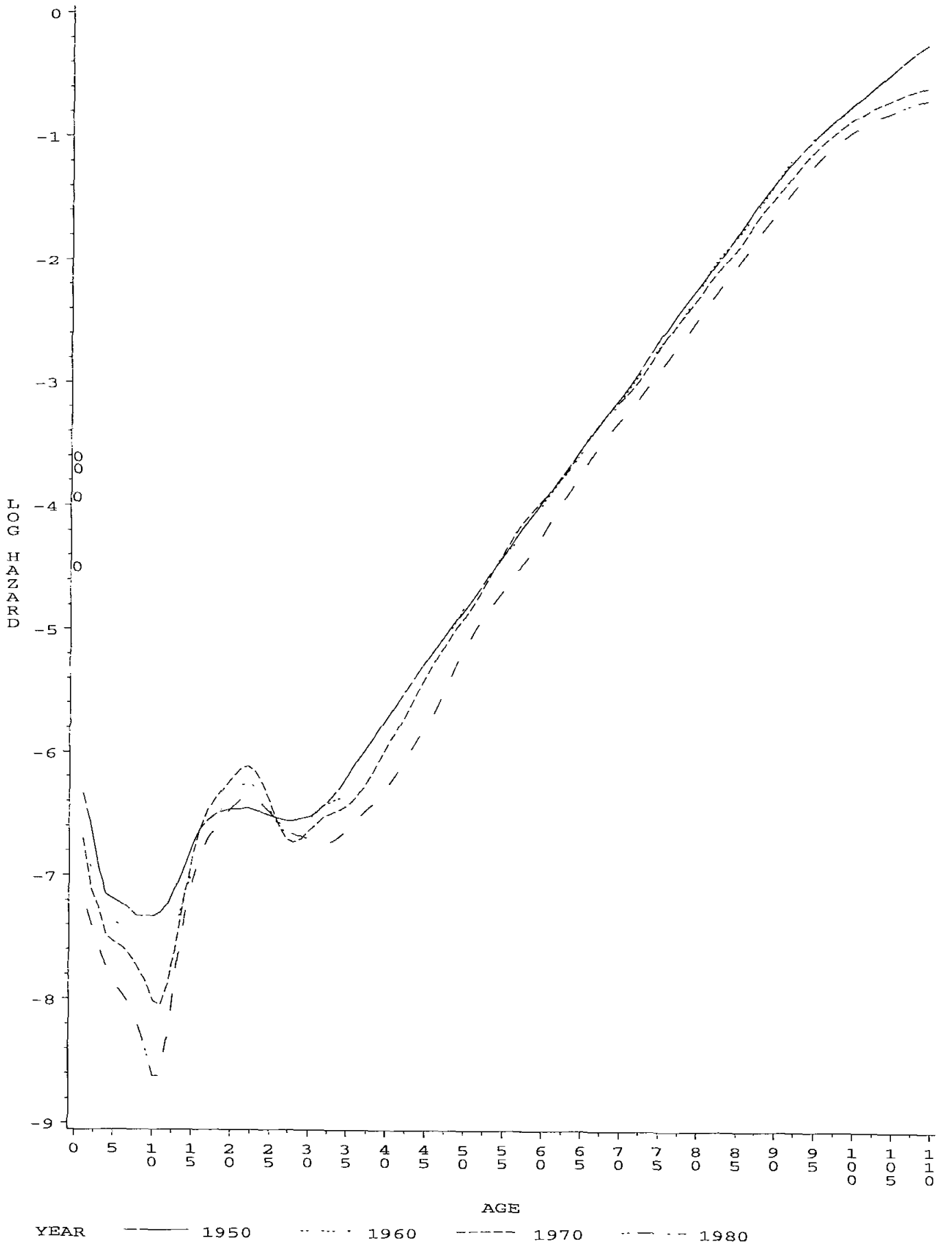

Log Hazard Rates
Minnesota White Male
LOG HAZARD
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
AGE
YEAR 1950 1960 1970 1980

Figure 3a

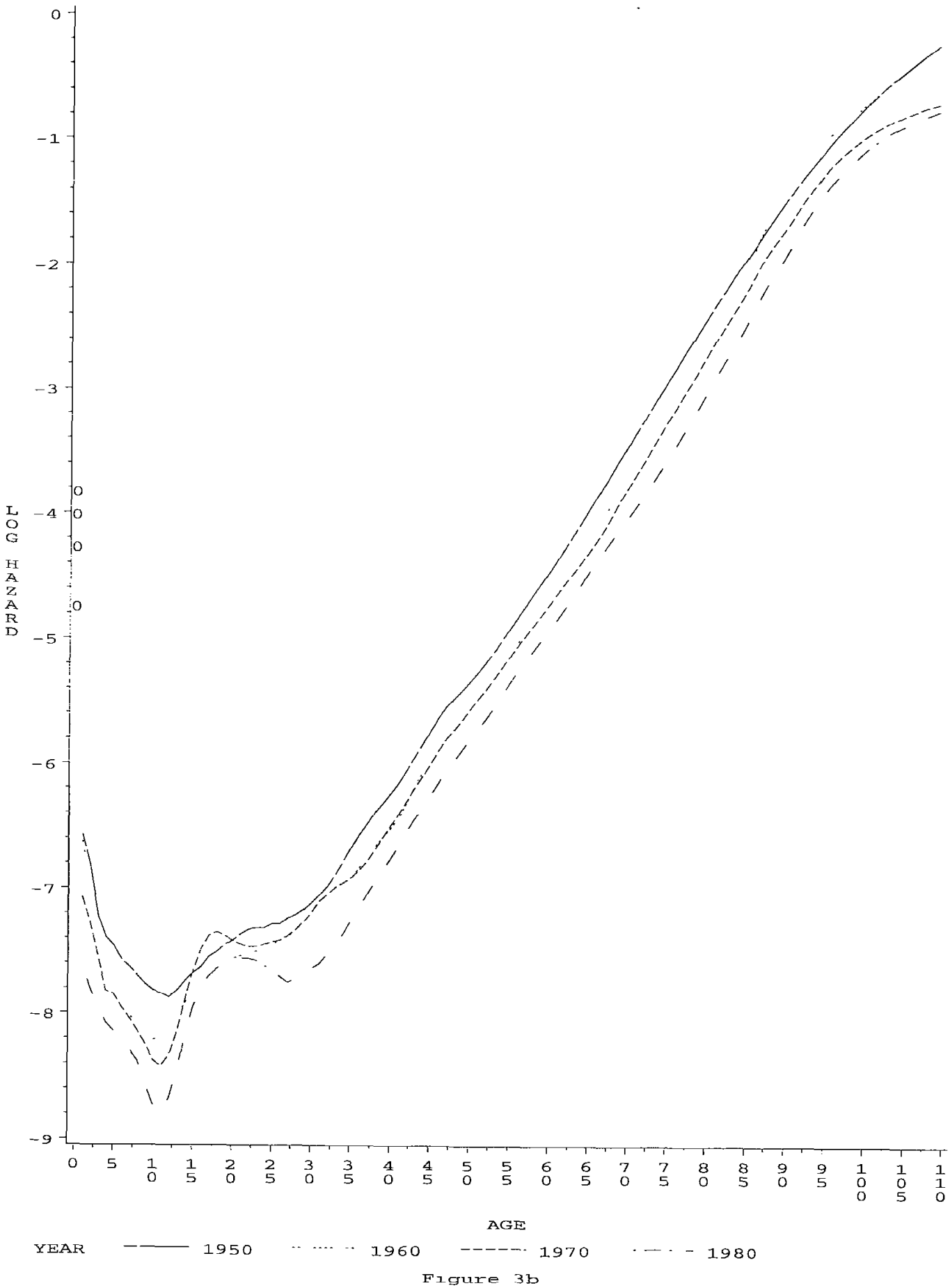
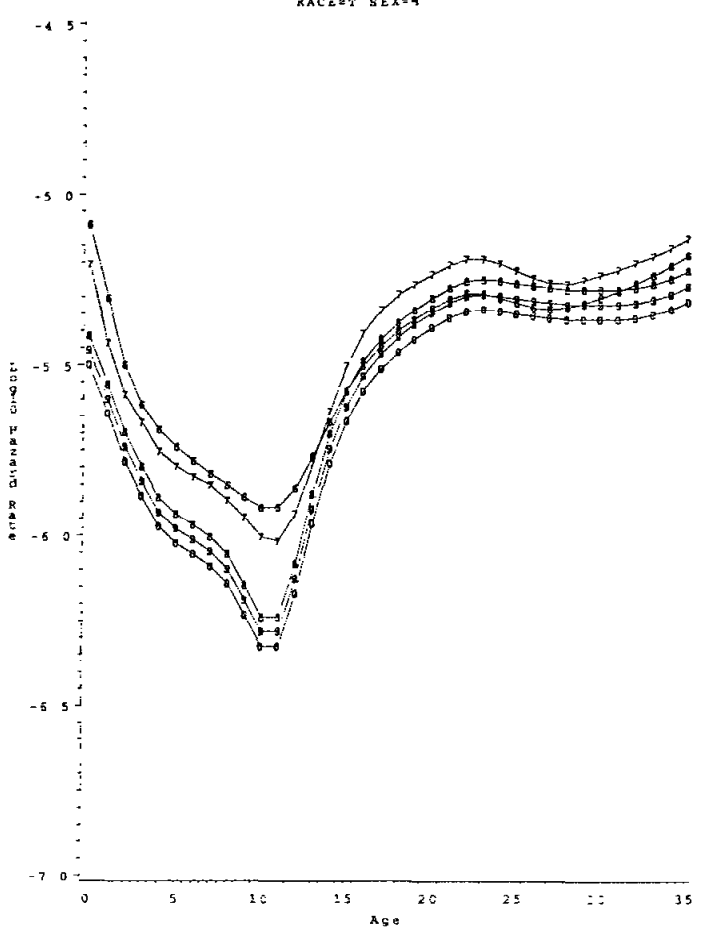
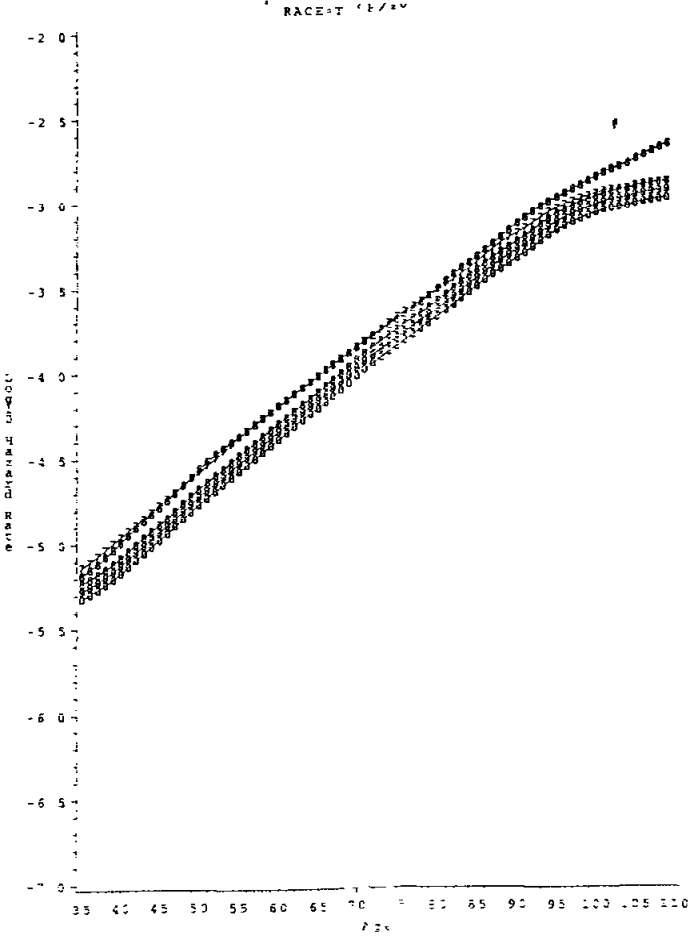
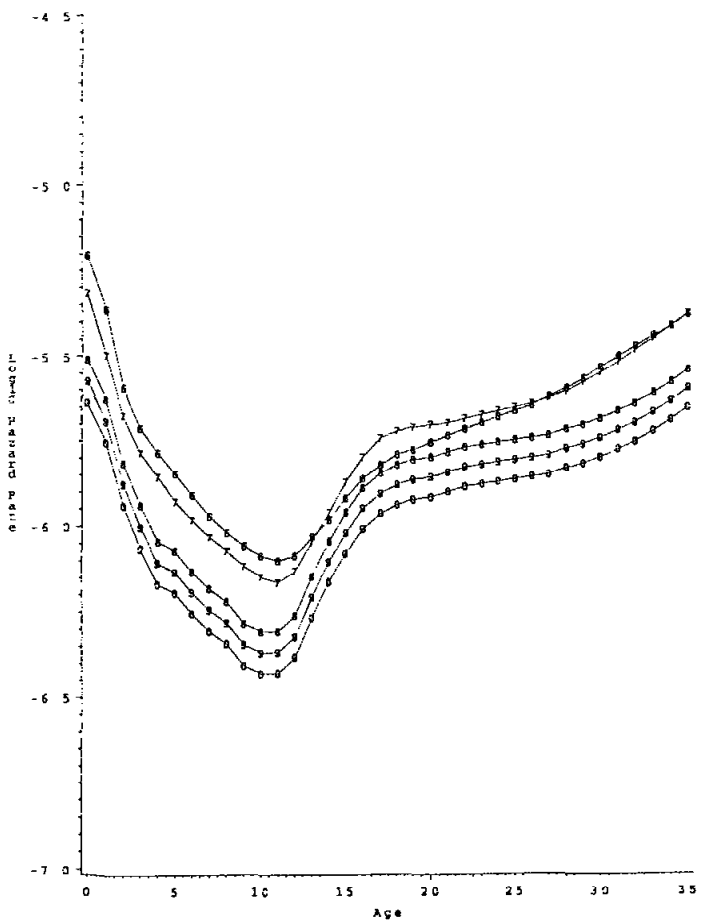

Log Hazard Rates
Minnesota White Female
LOG HAZARD
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
AGE
YEAR 1950 1960 1970 1980

Figure 3b

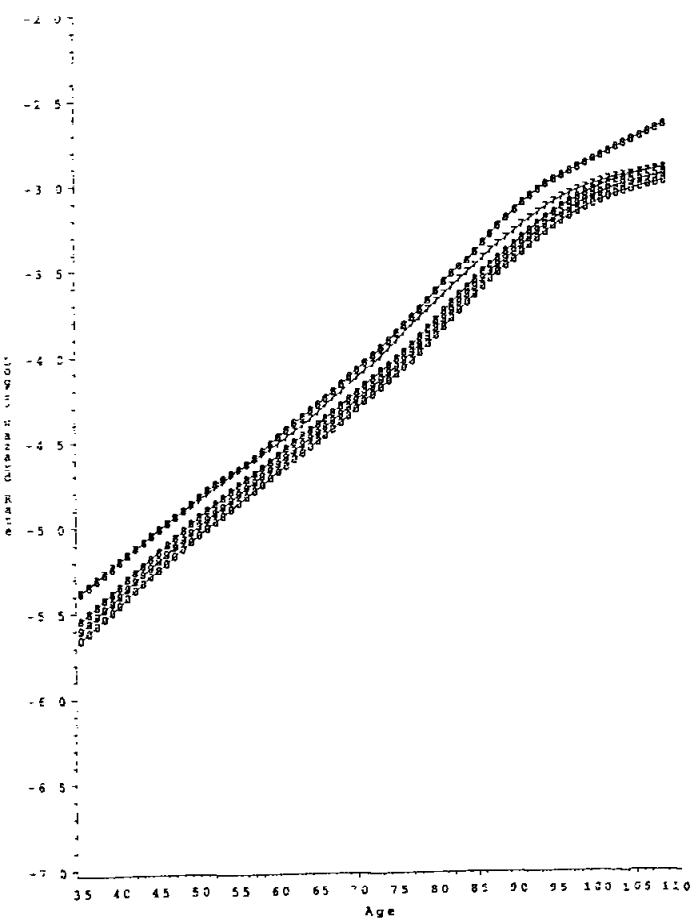




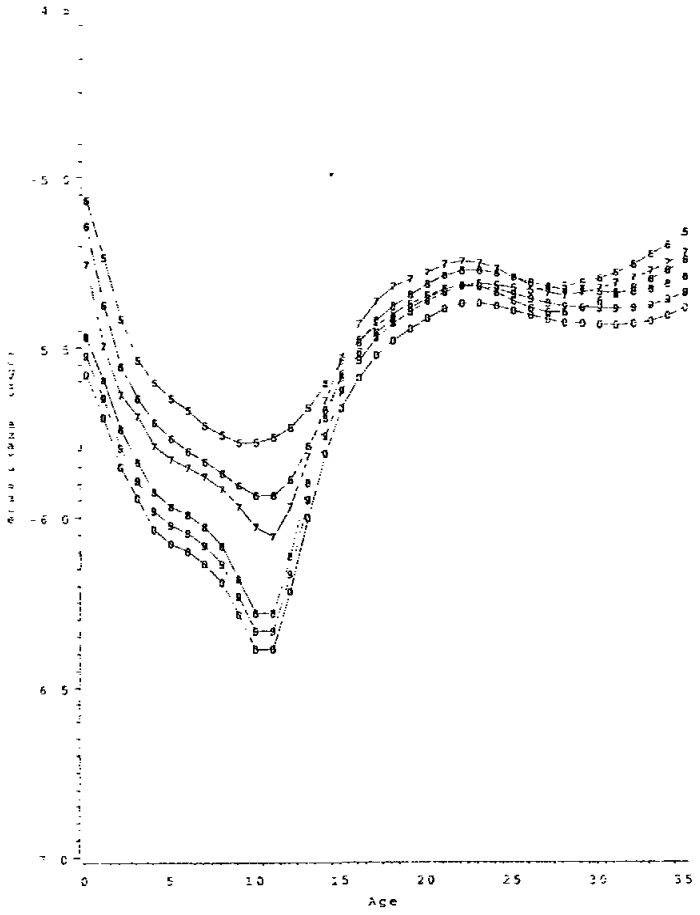
US Population Age 0 to 35



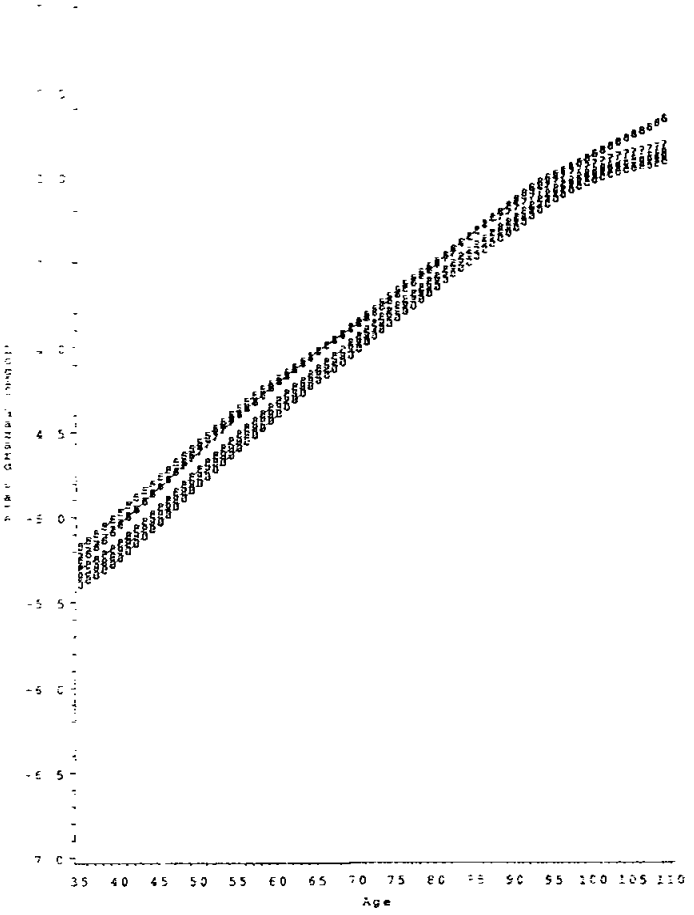
US Population Age 35 to 109



US Population Age 0 to 35

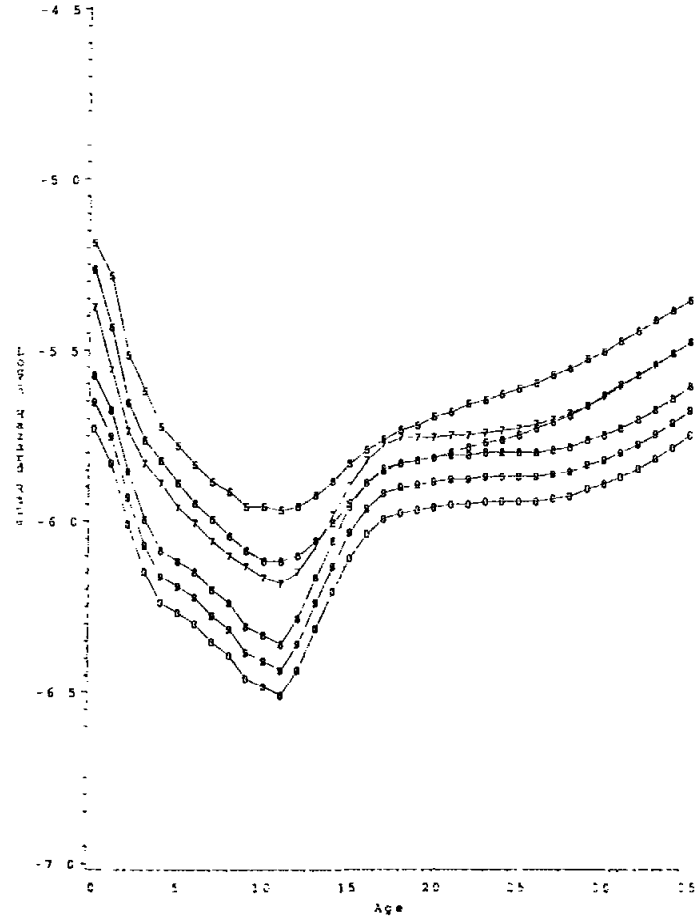


US Population Age 35 to 109



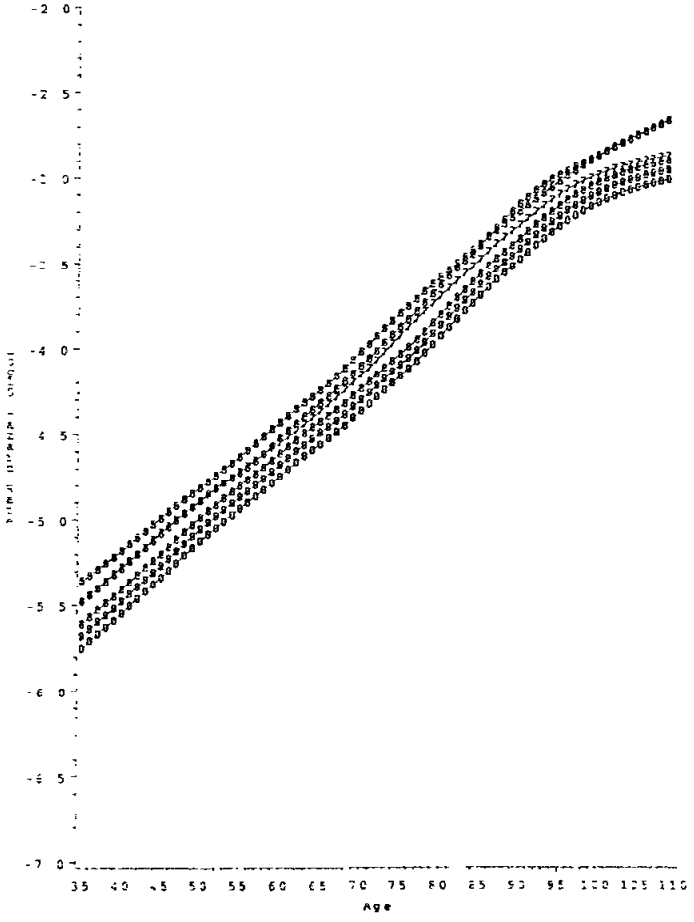
US Population Age 0 to 35

RACE=W SEX=F

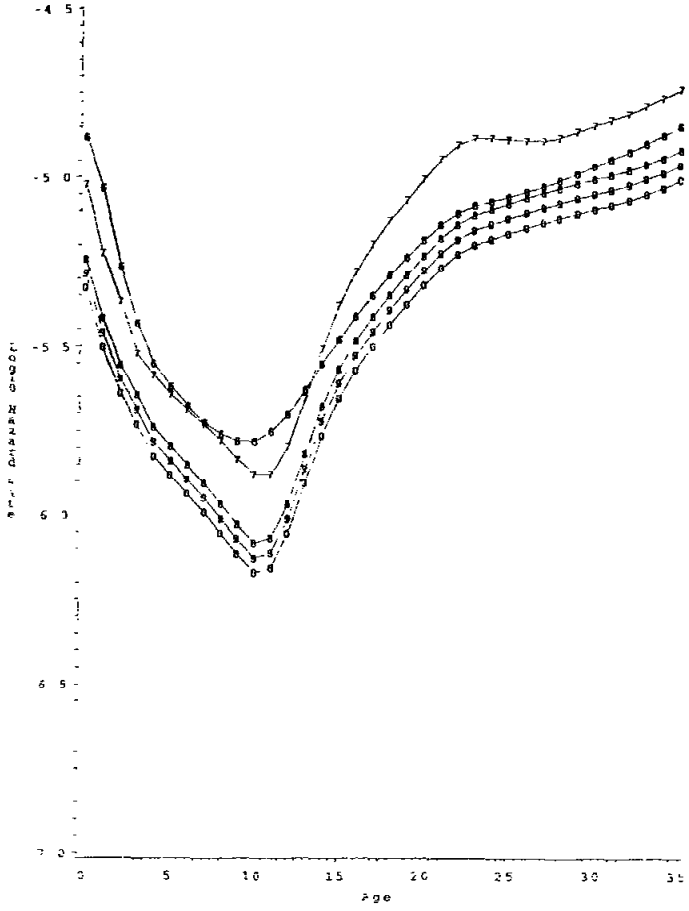


US Population Age 35 to 109

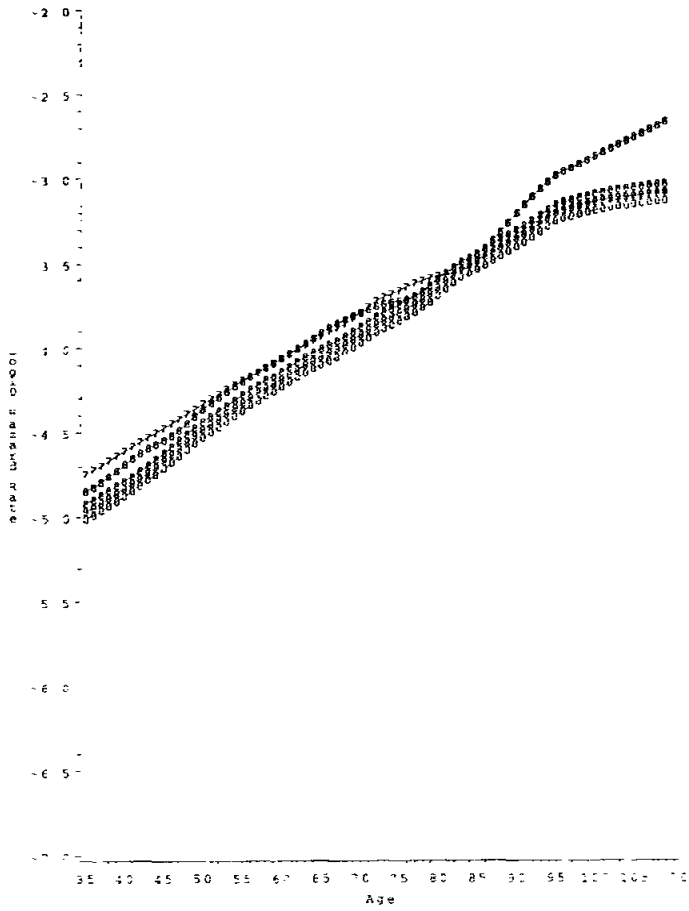
RACE=W SEX=F



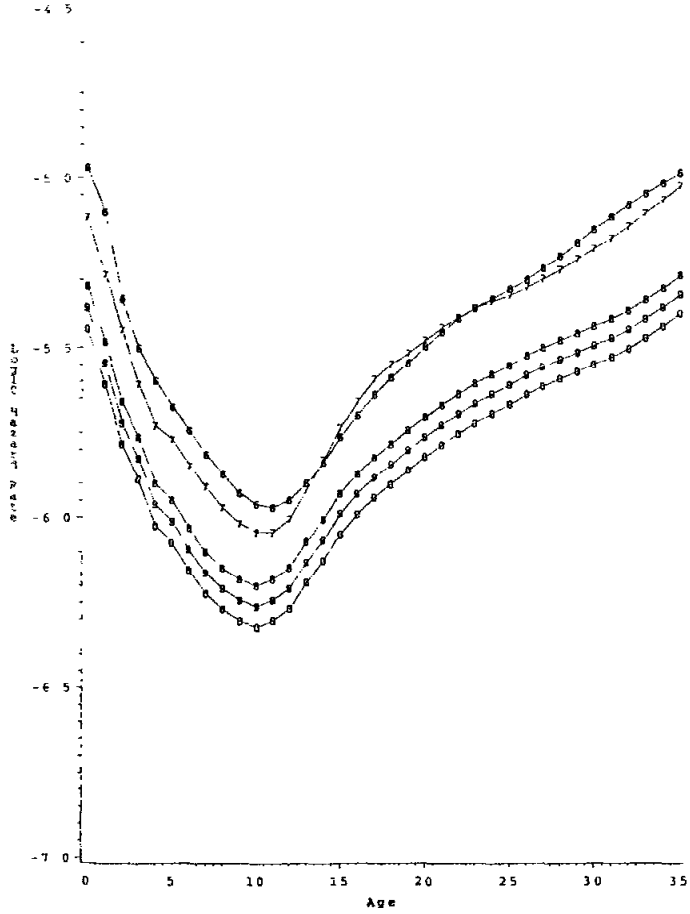
US Population Age 0 to 35
RACE=NW SEX=M



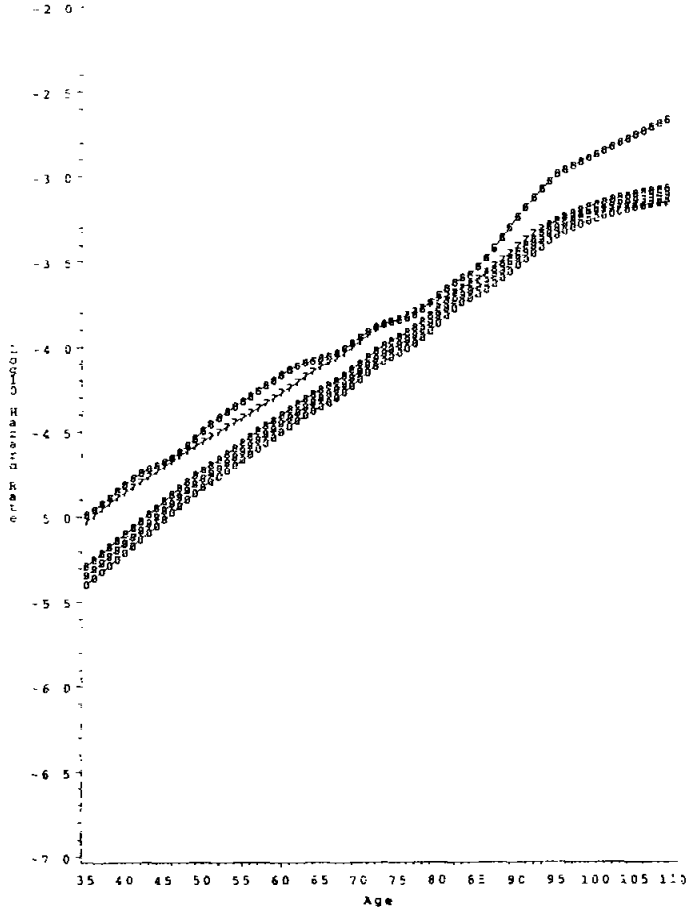
US Population Age 35 to 109
RACE=NW SEX=M

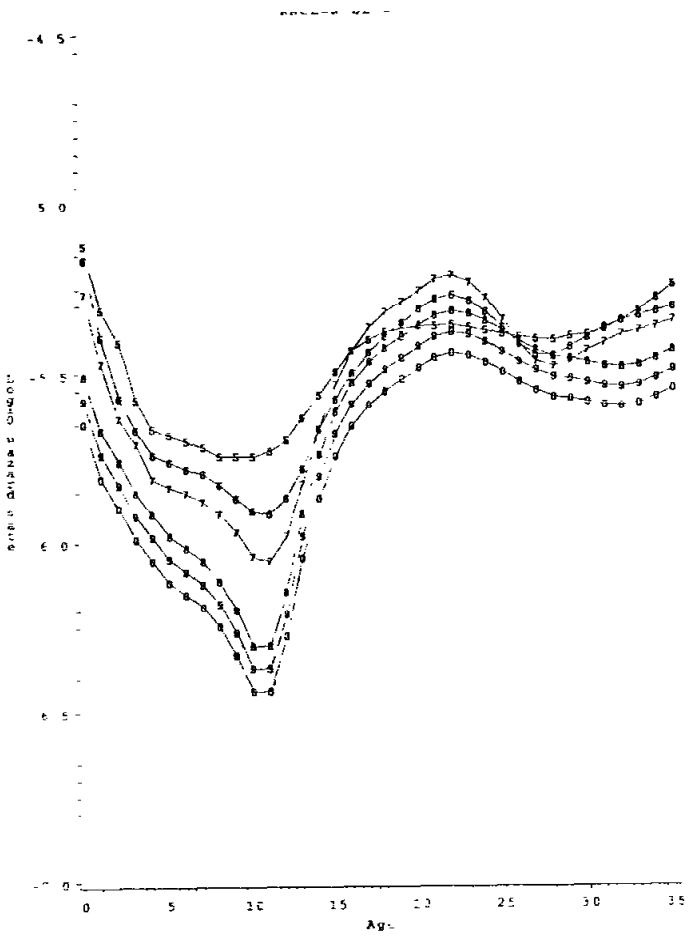


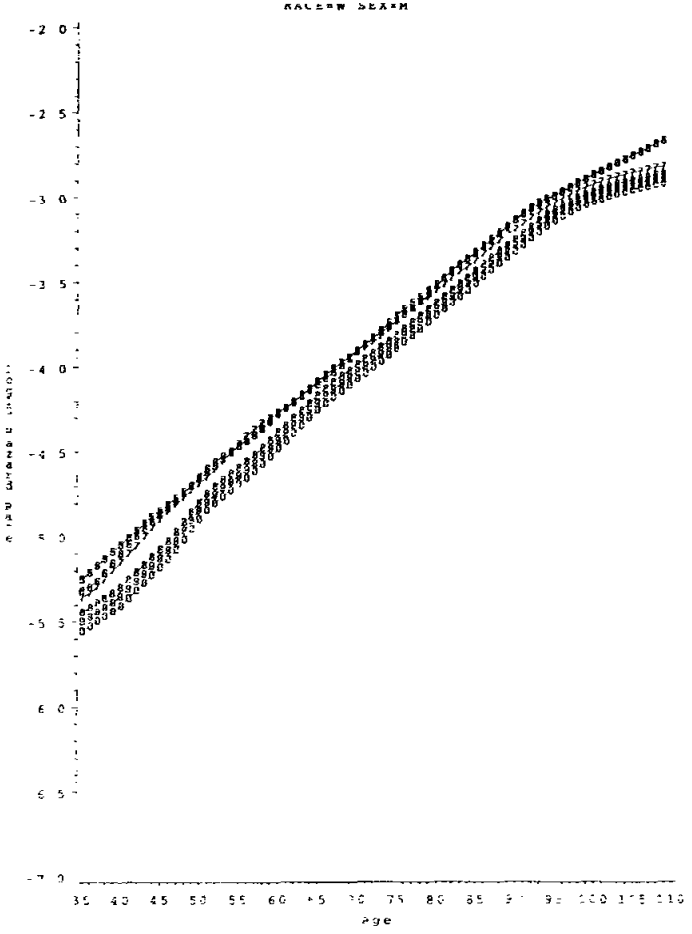
US Population Age 0 to 35
RACE=NW SEX=F



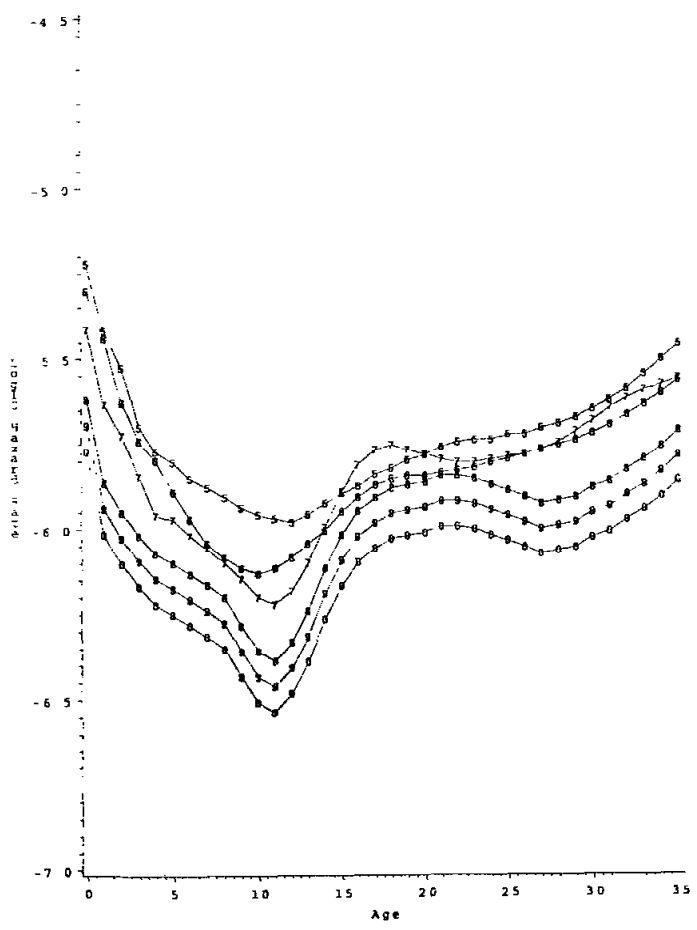
US Population Age 35 to 109
RACE=NW SEX=F



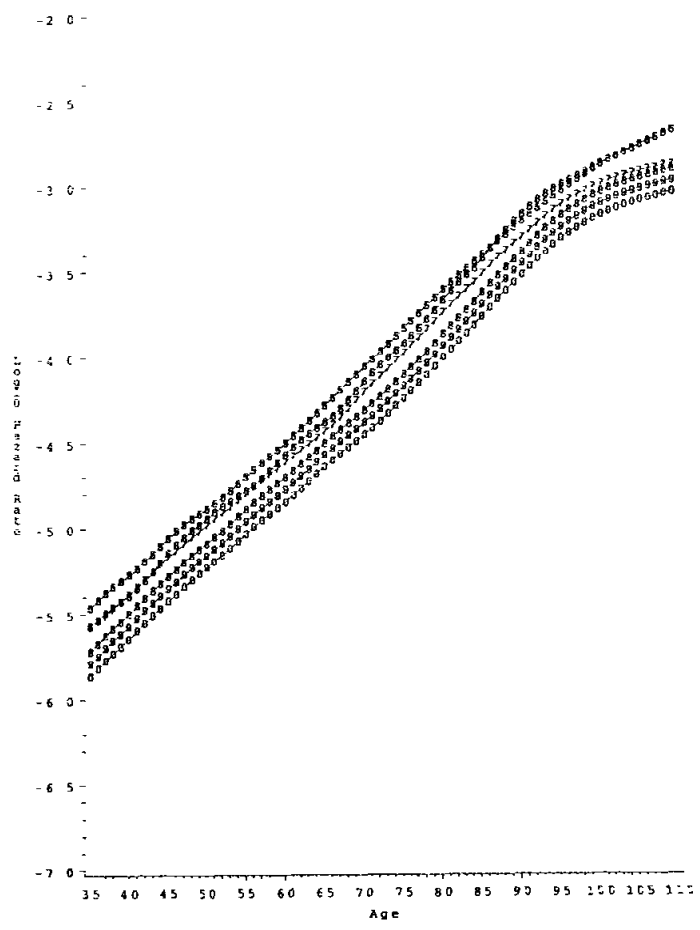


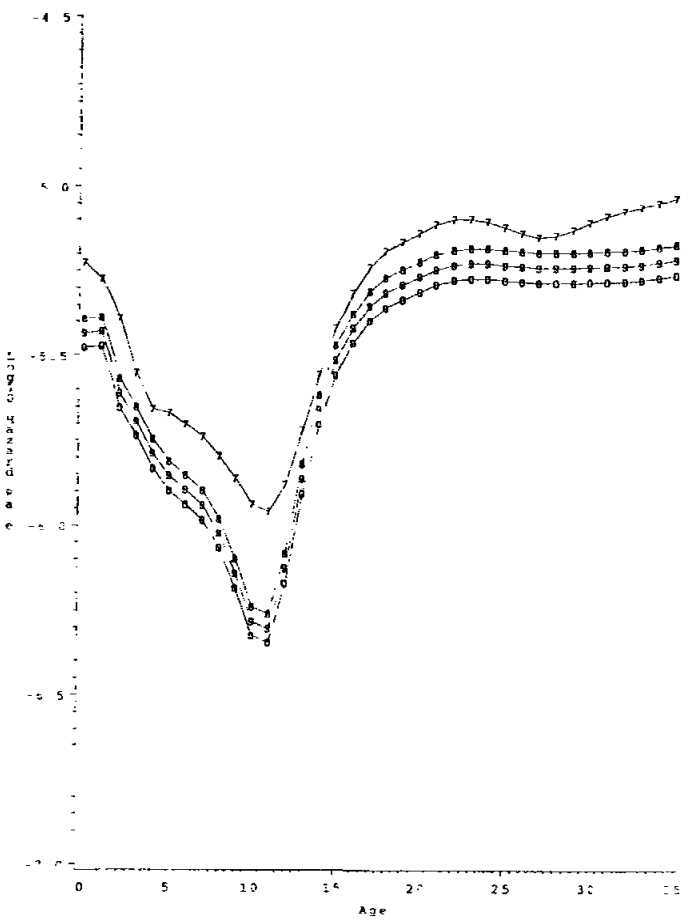


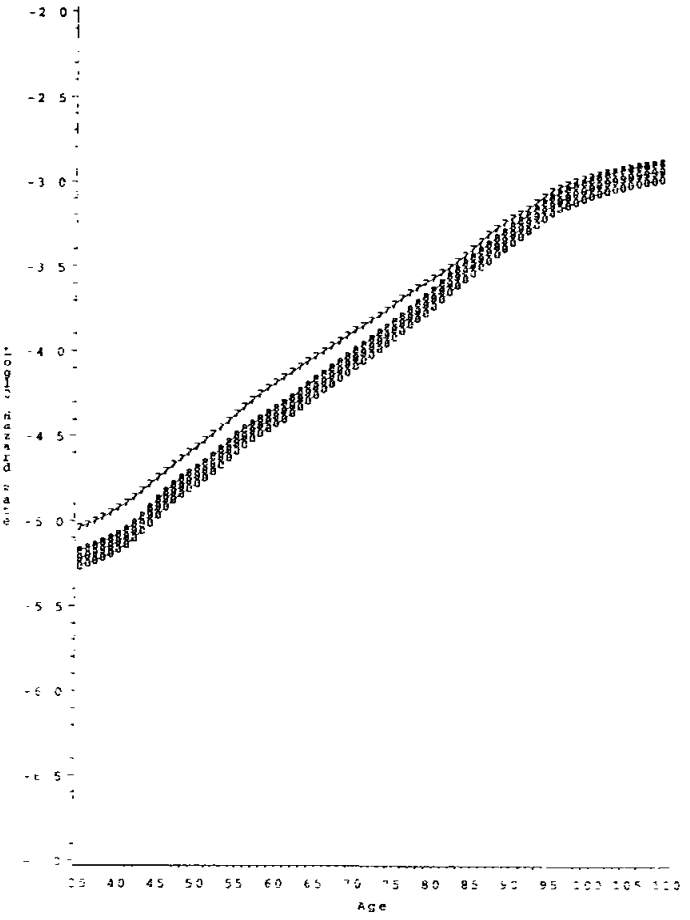
MN Population Age 0 to 35



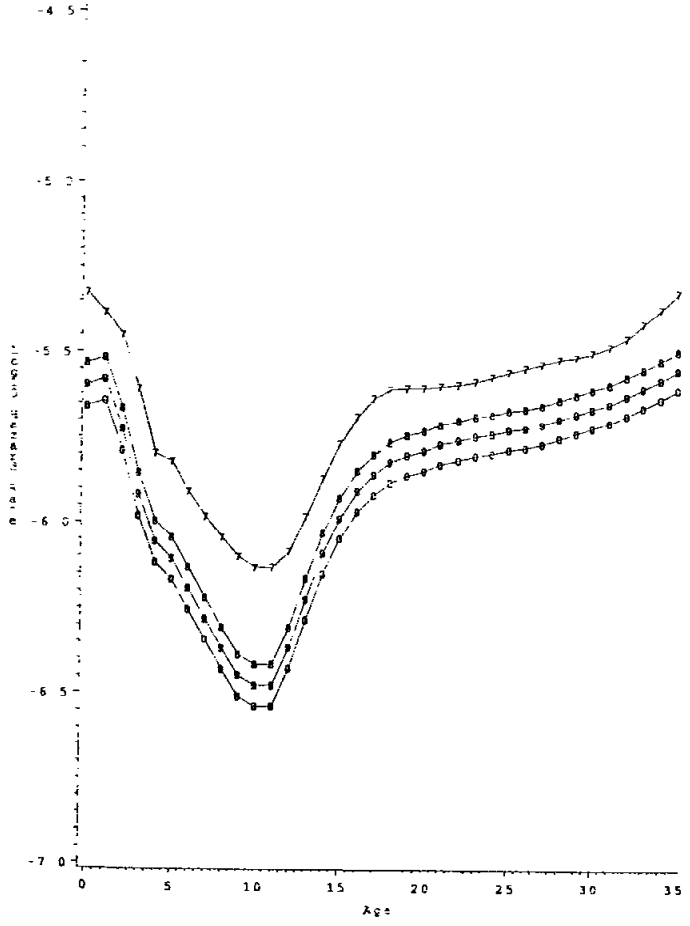
MN Population Age 35 to 109



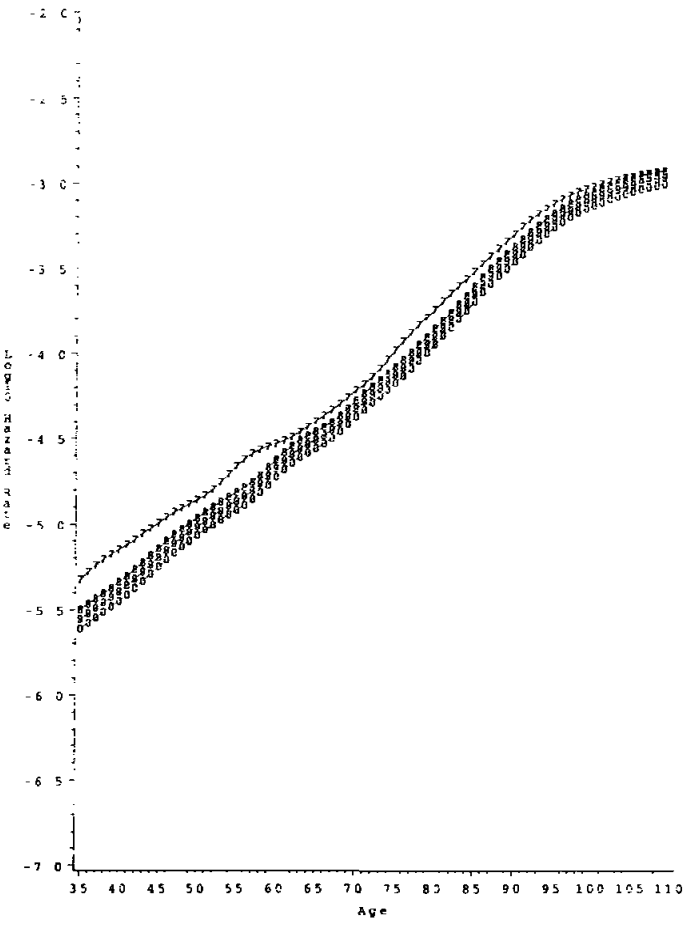


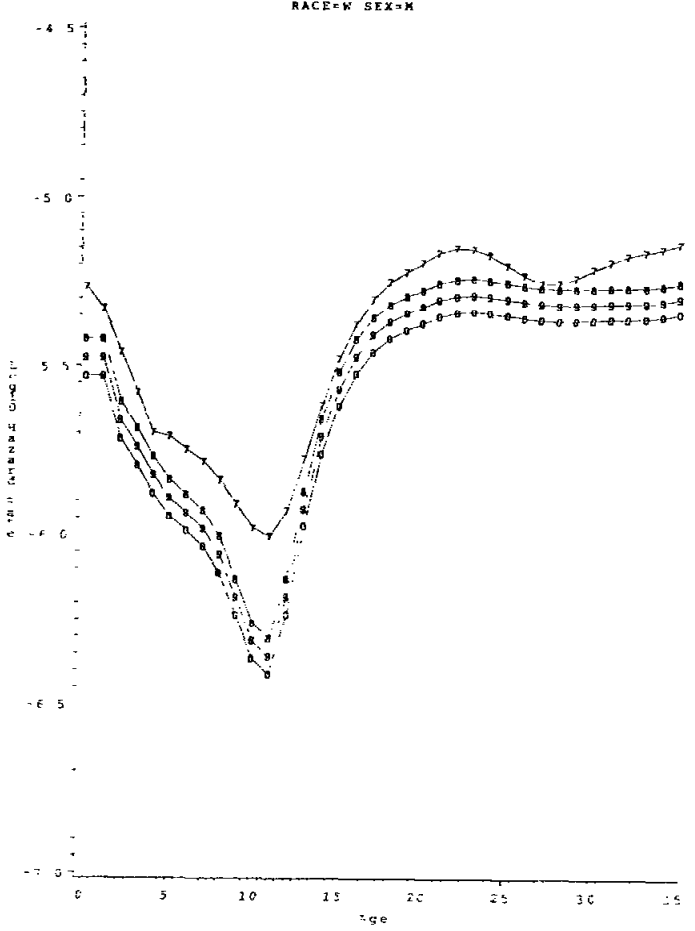


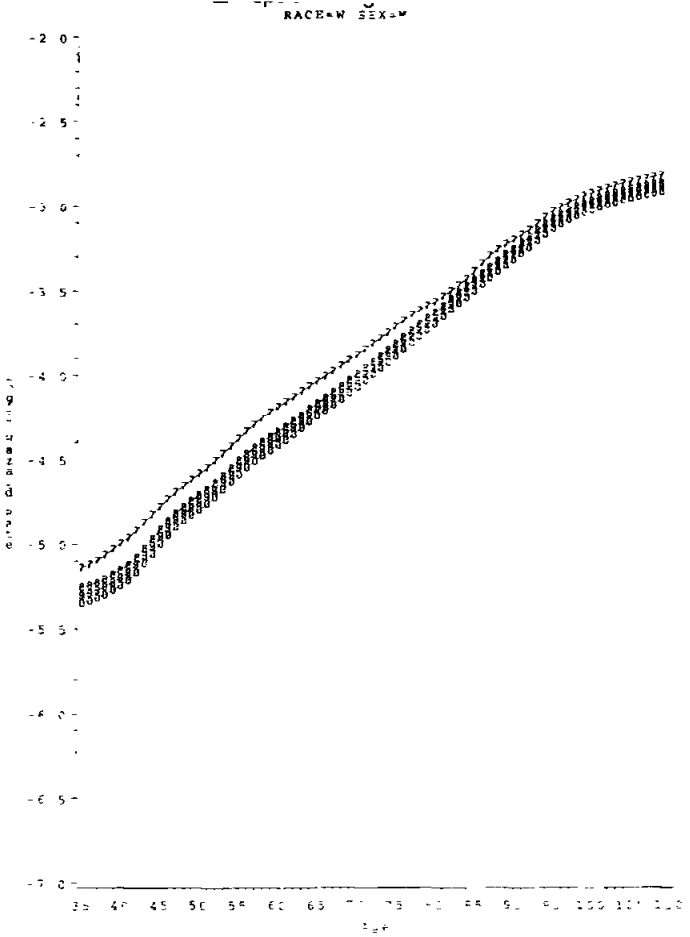
AZ Population Age 0 to 35



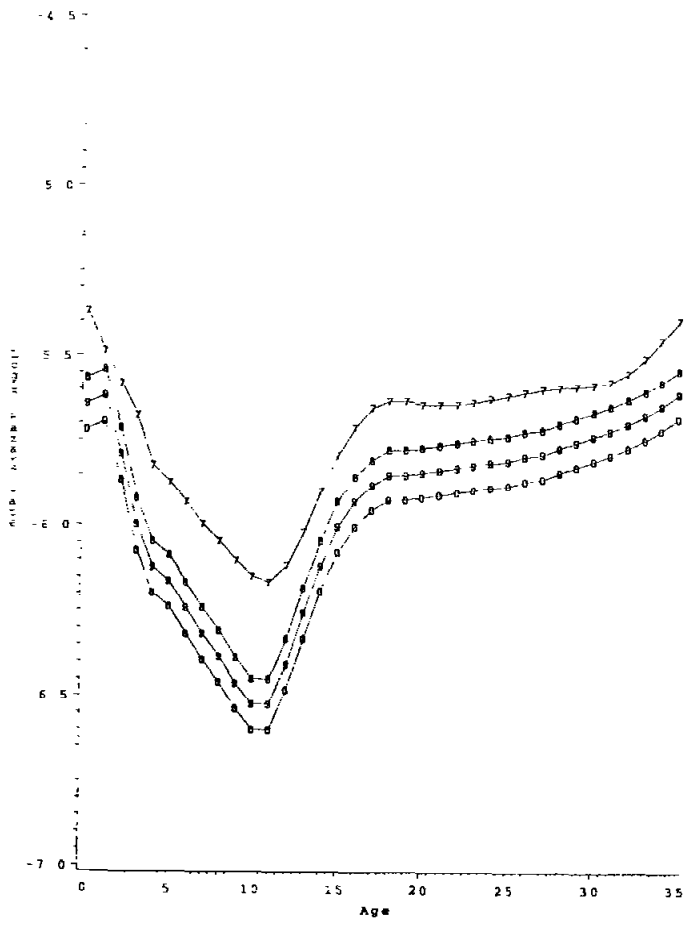
AZ Population Age 35 to 109



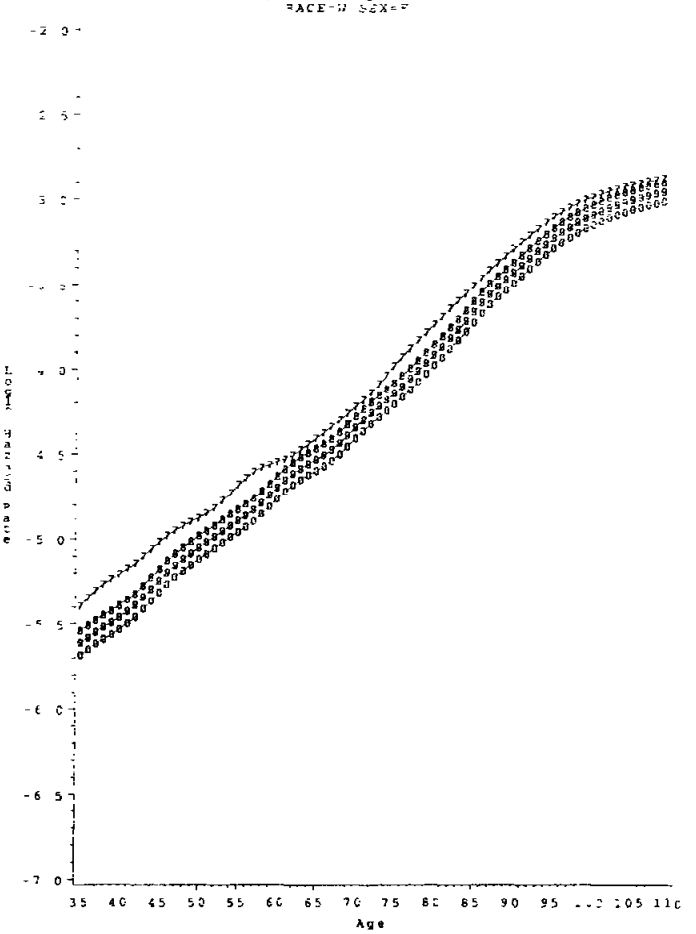


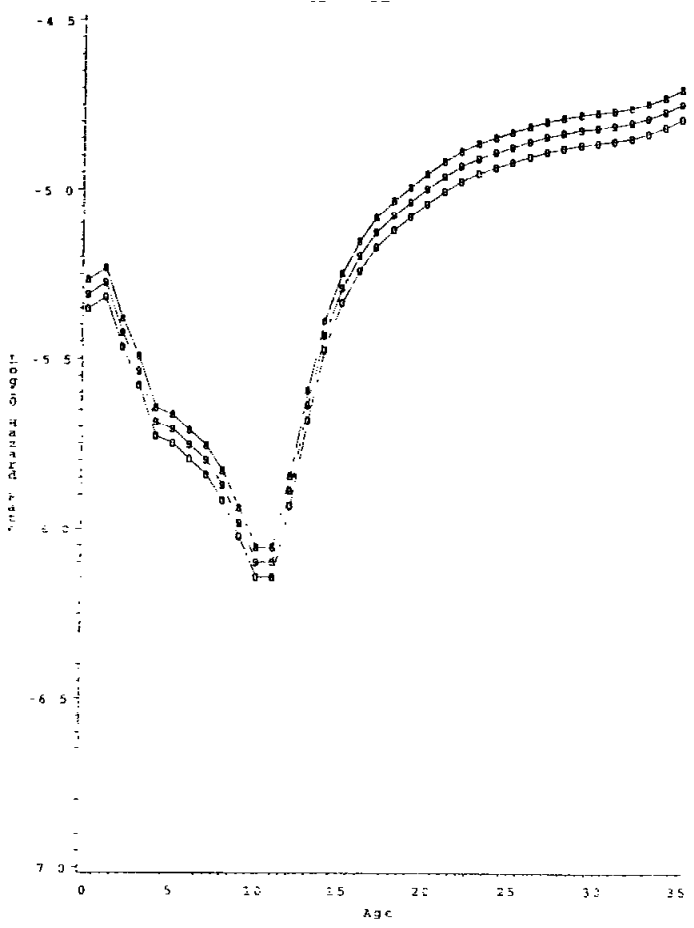


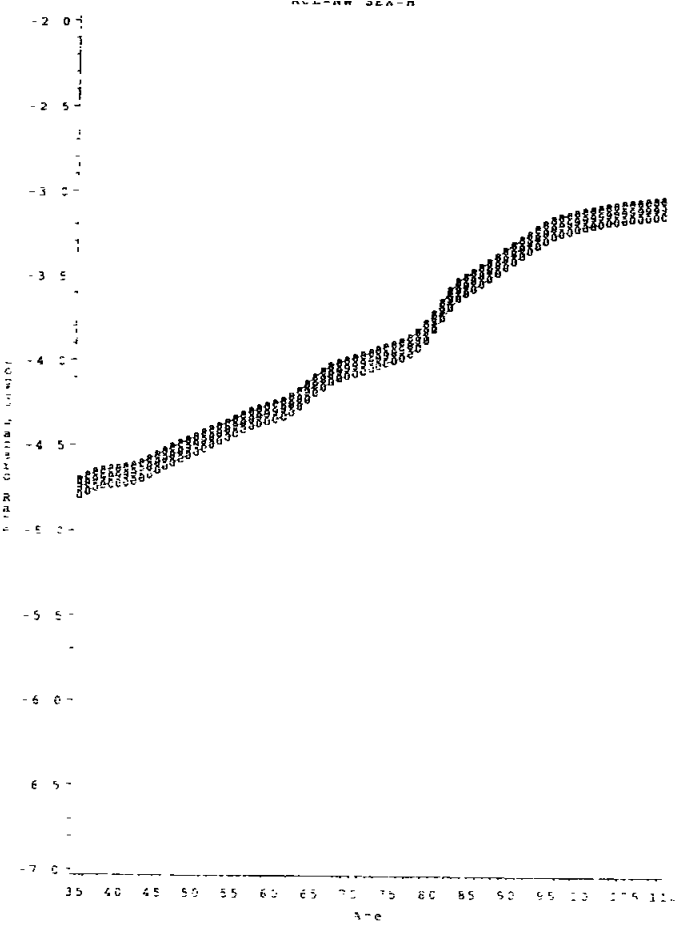
AZ Population Age 0 to 35



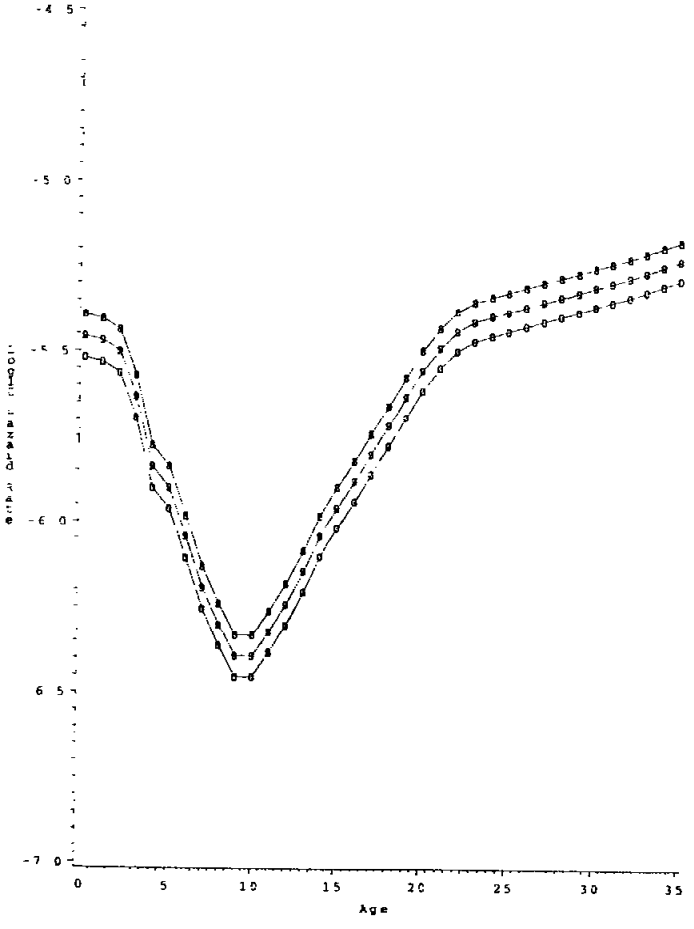
AZ Population Age 35 to 109



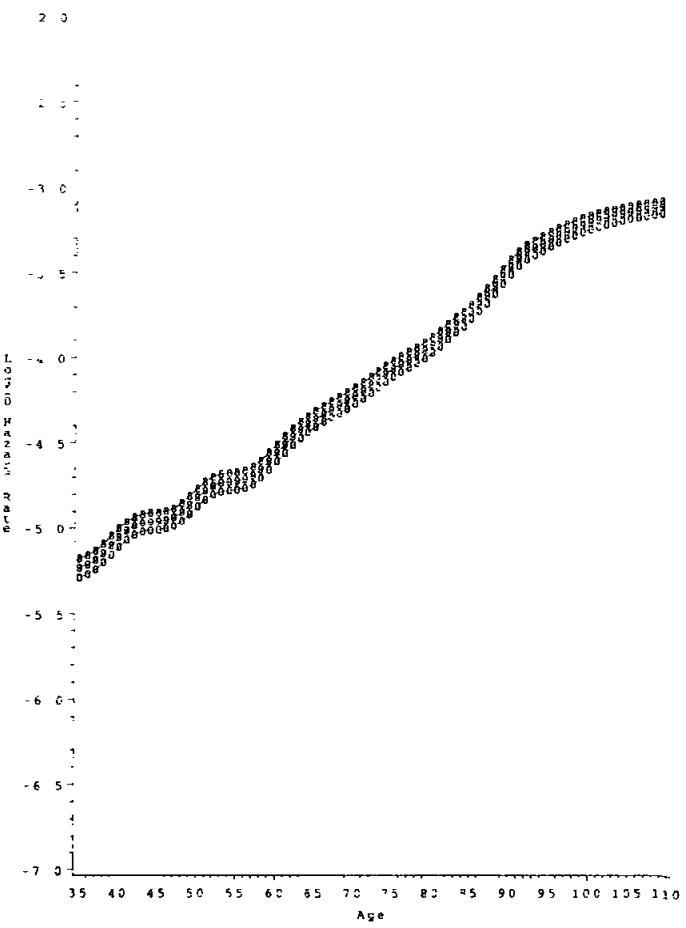


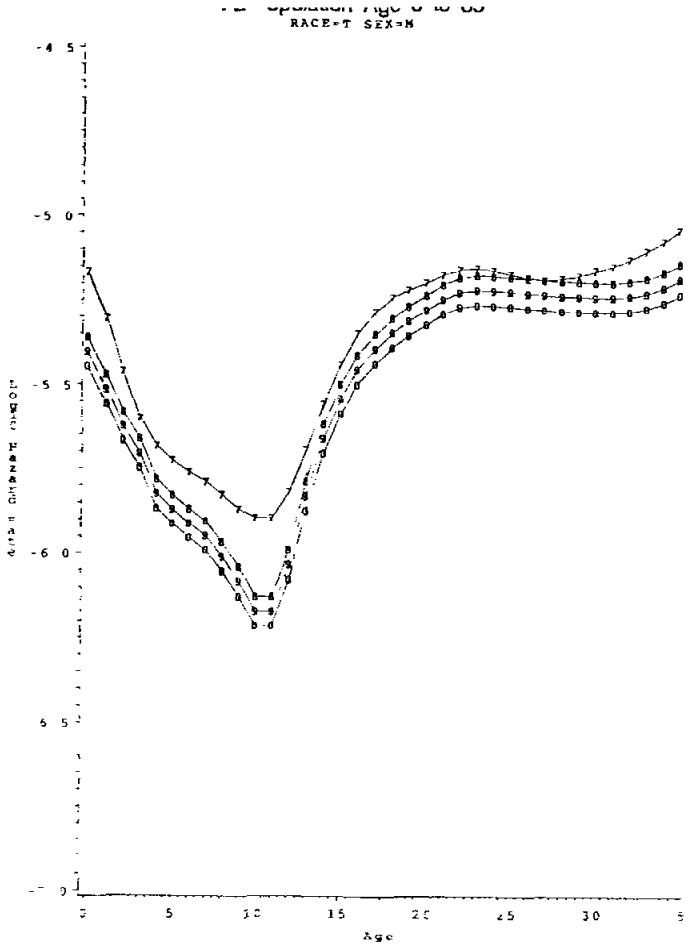


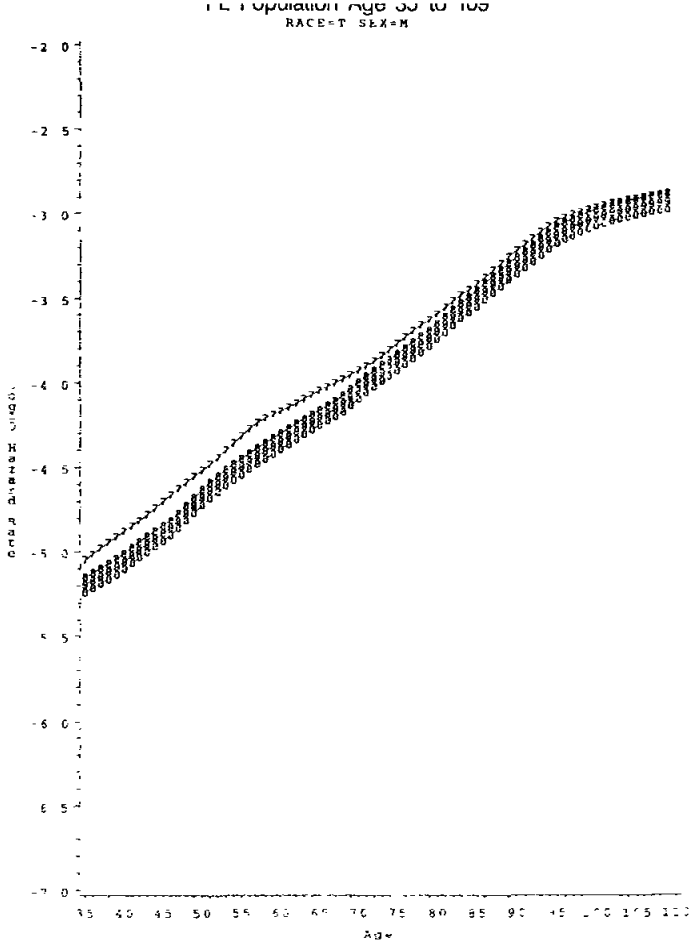
AZ Population Age 0 to 35



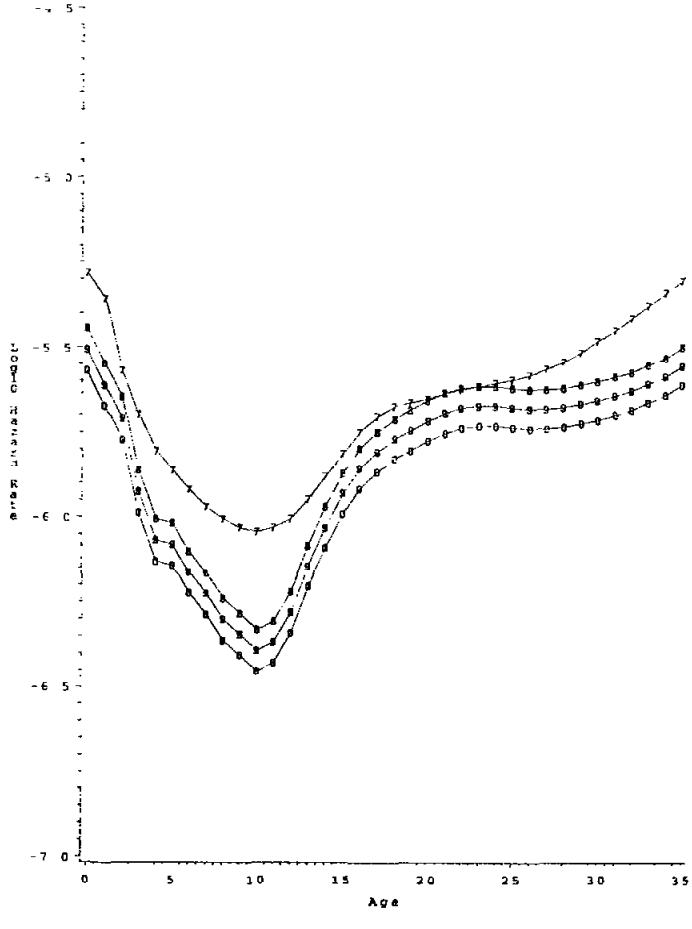
AZ Population Age 35 to 109



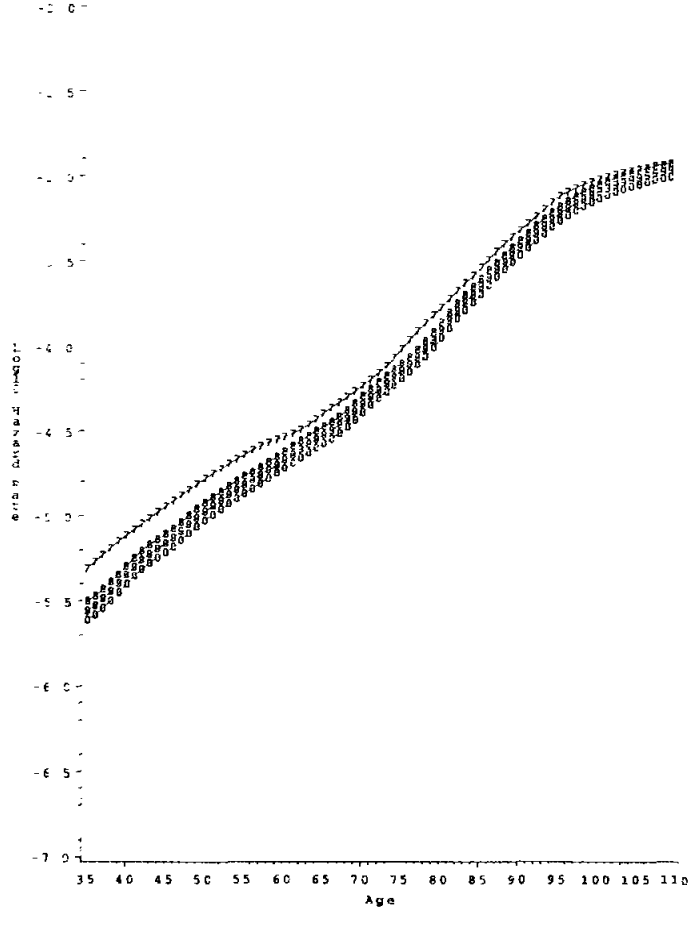




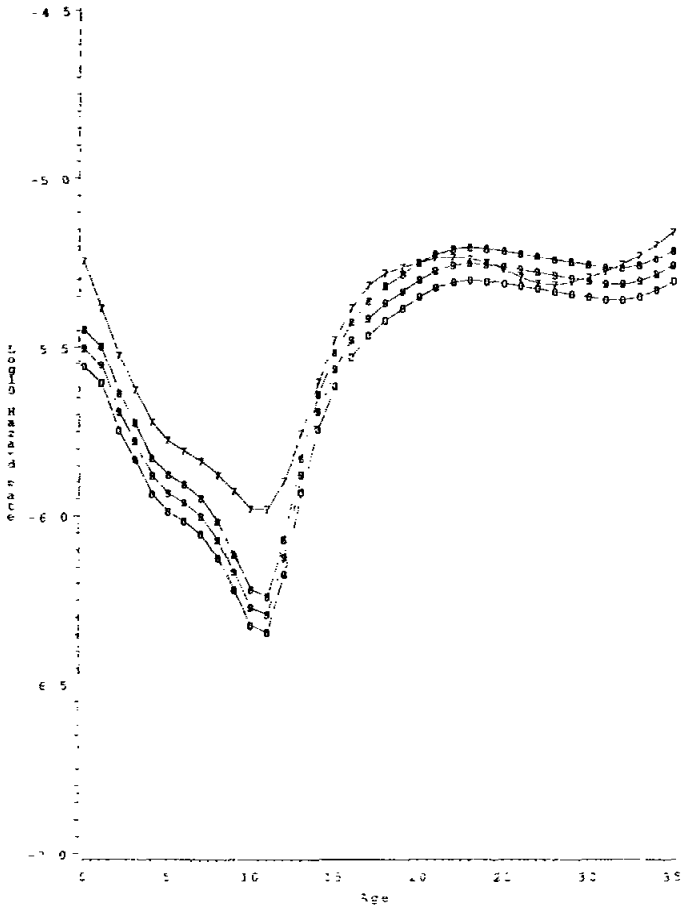
FL Population Age 0 to 35



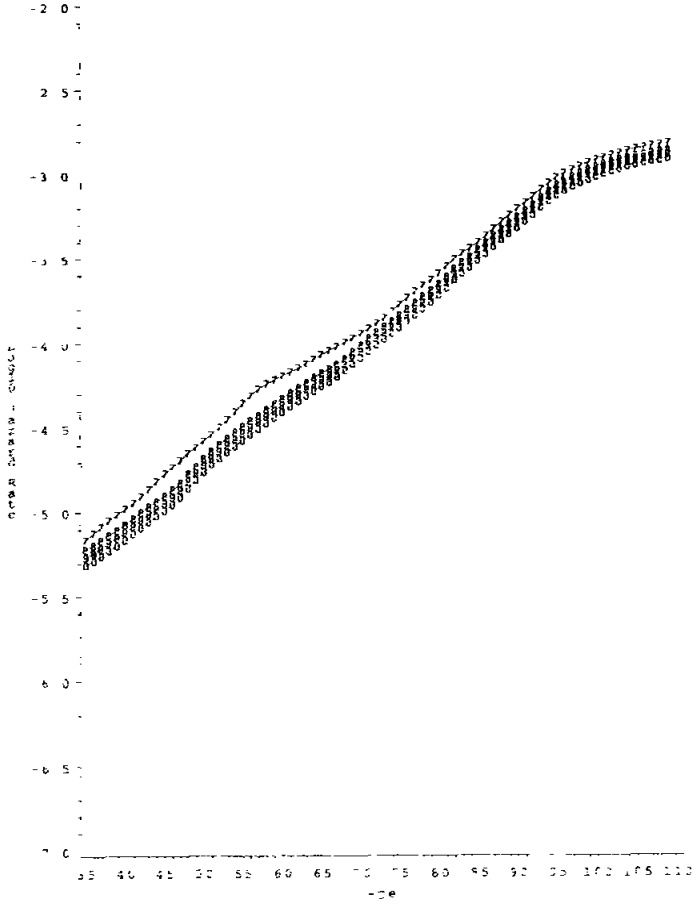
FL Population Age 35 to 109



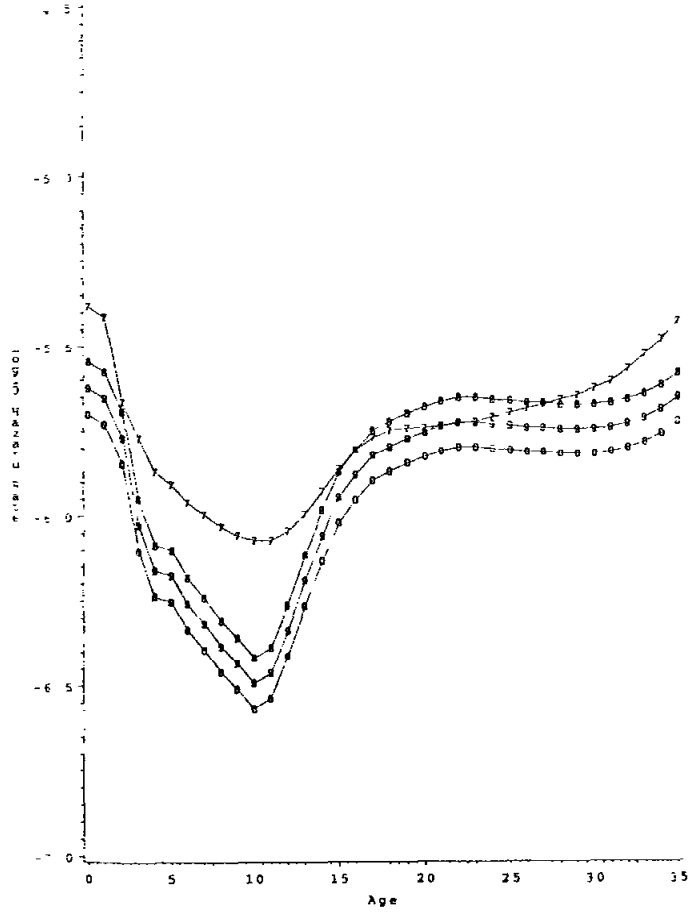
FL Population Age 0 to 35



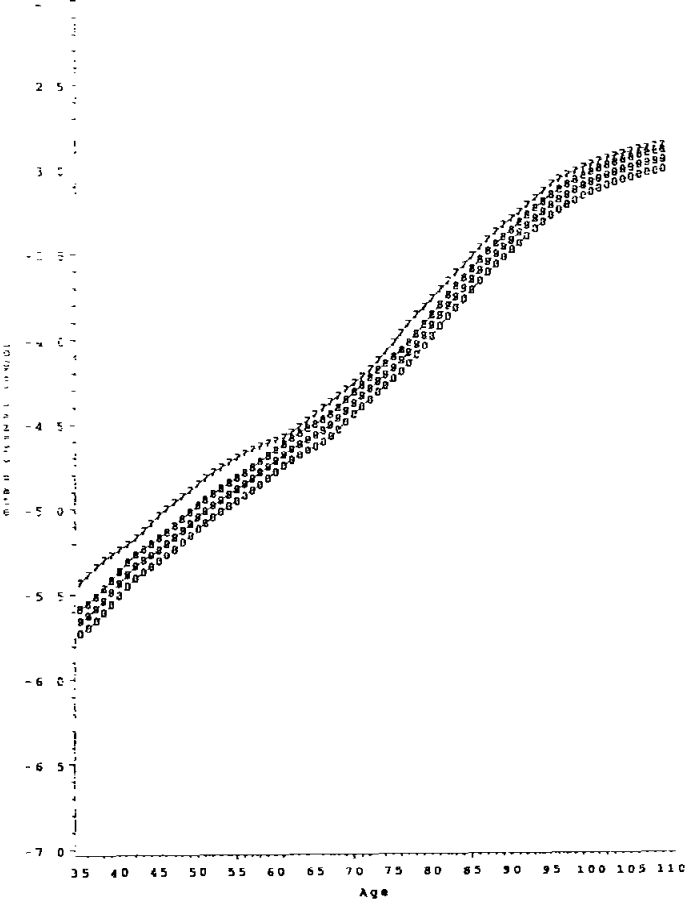
FL Population Age 35 to 109

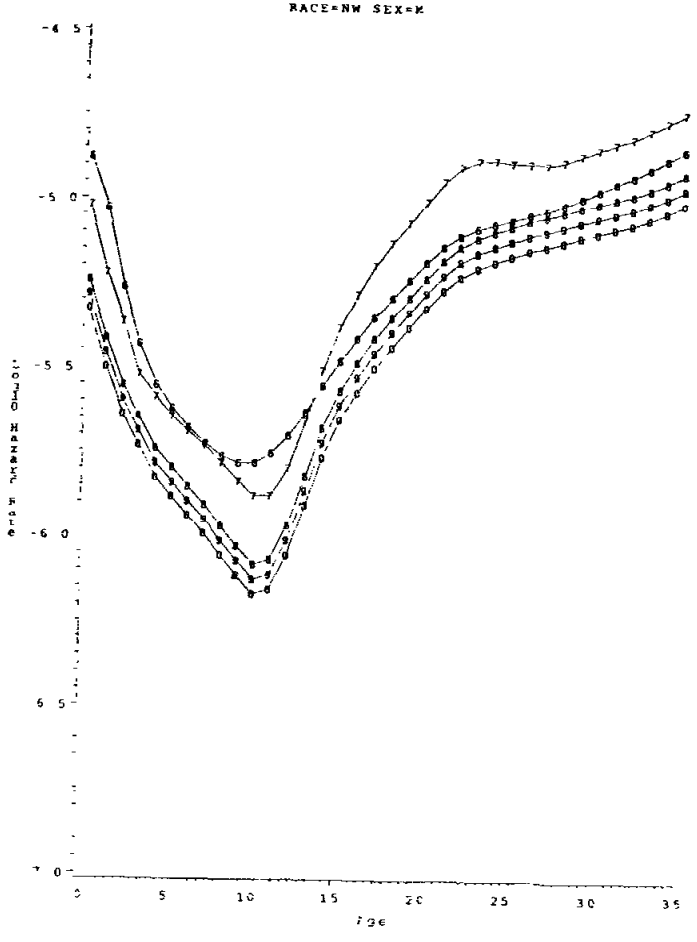


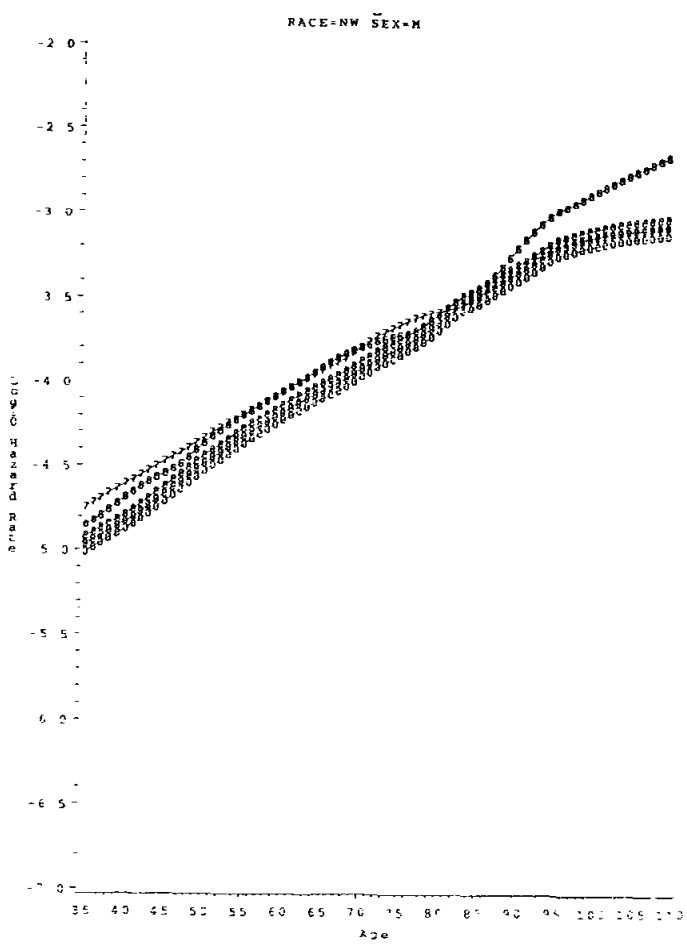
FL Population Age 0 to 35



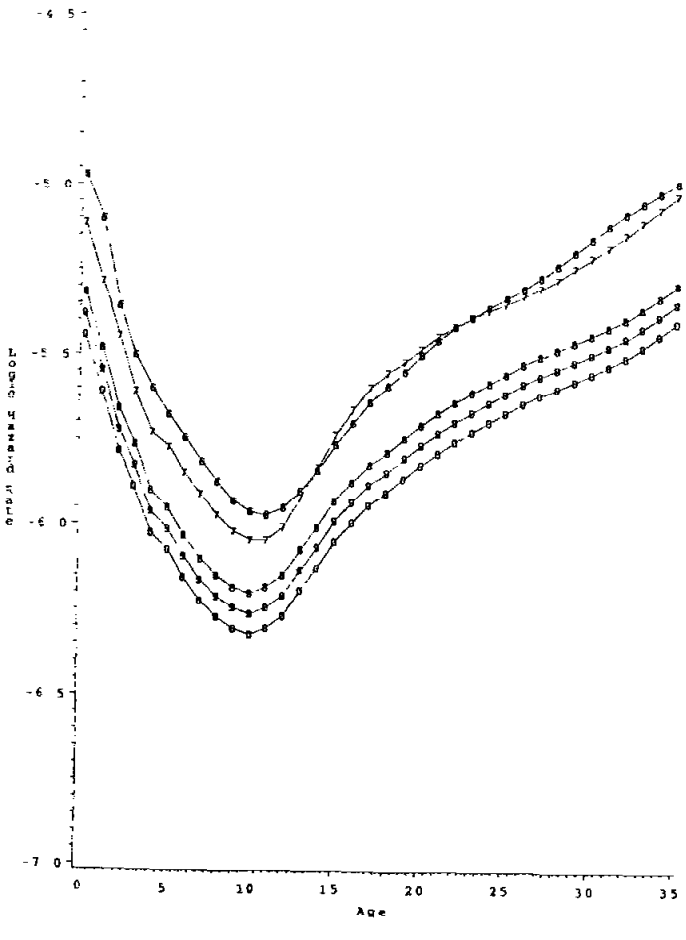
FL Population Age 35 to 109



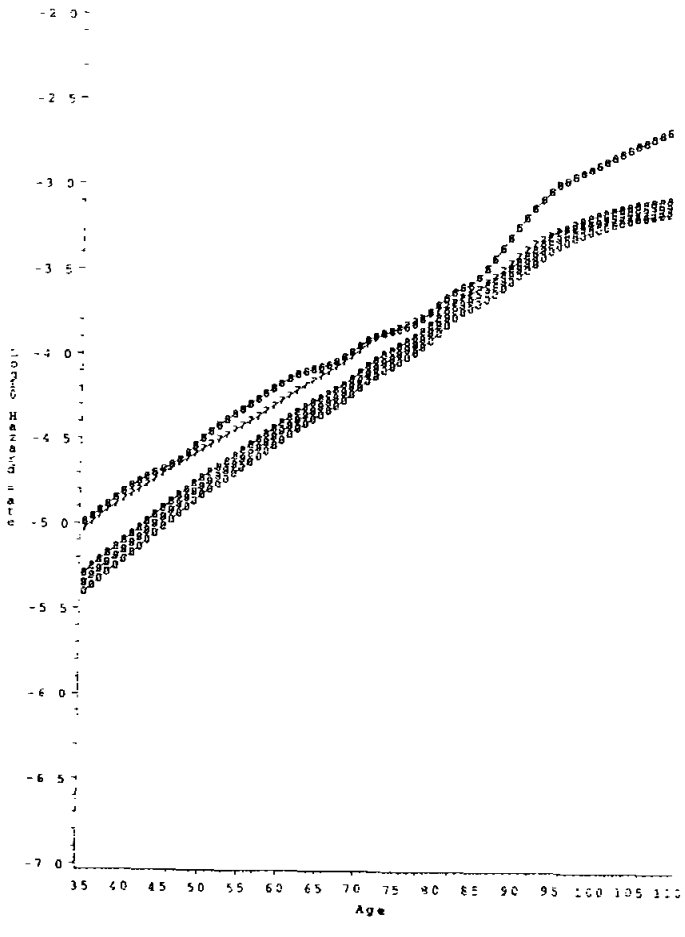




US Population Age 0 to 35

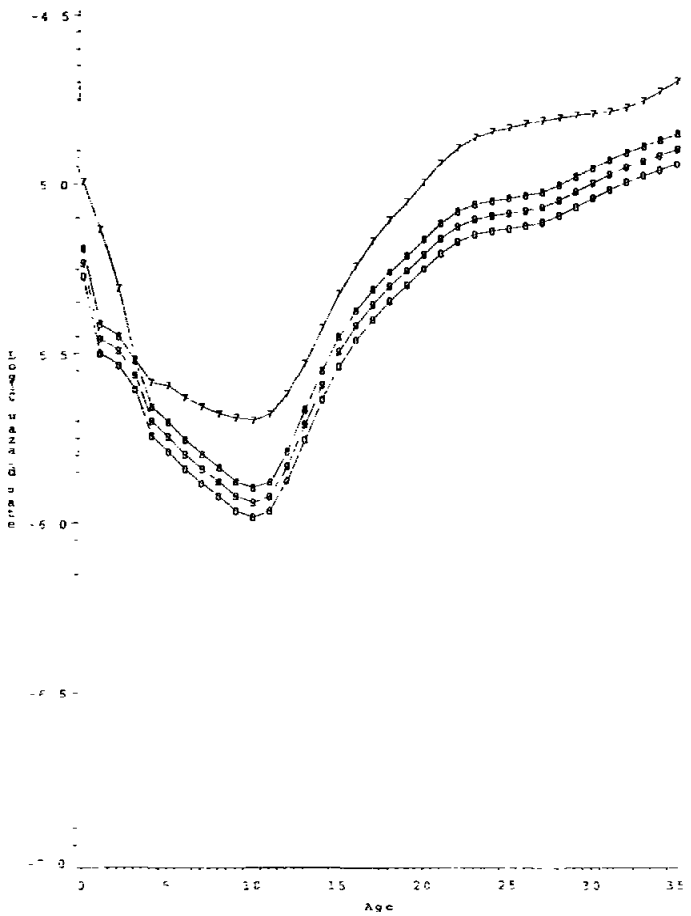


US Population Age 35 to 109



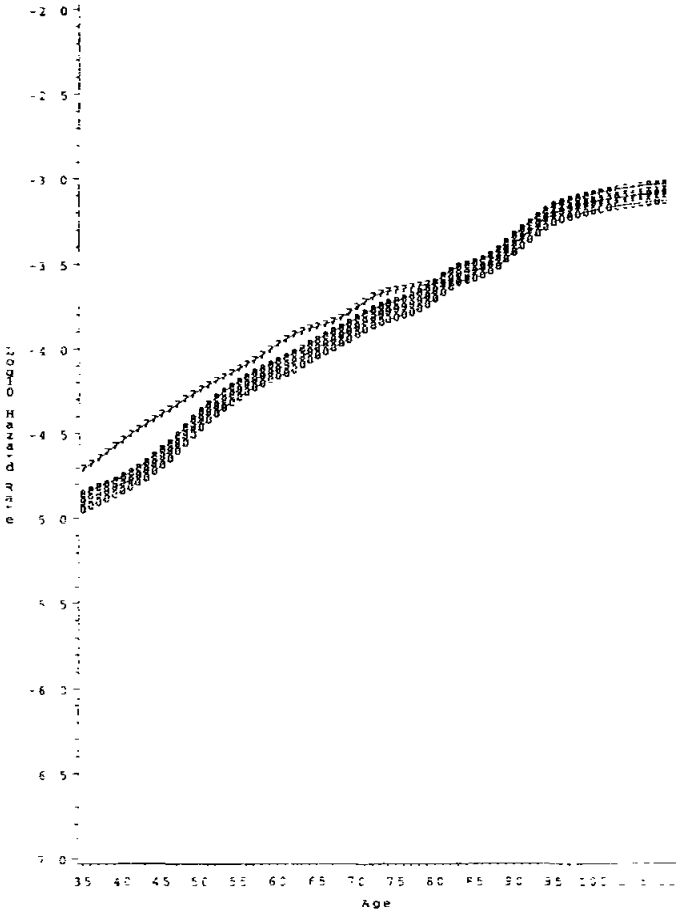
FL Population Age 0 to 35

RACE=NW SEX=M



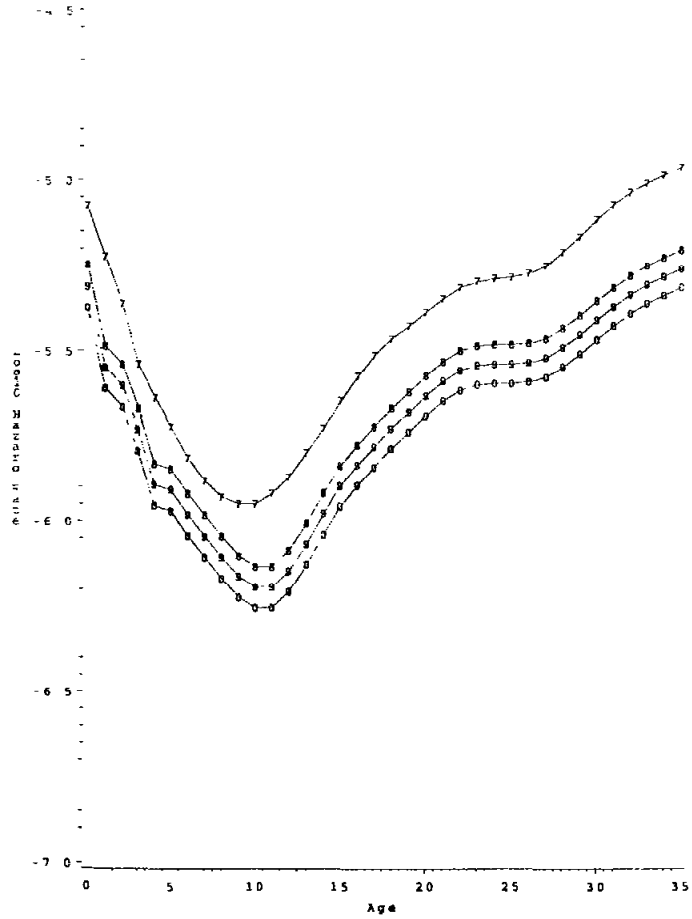
FL Population Age 35 to 109

RACE=NW SEX=M



FL Population Age 0 to 35

RACE=NW SEX=F



FL Population Age 35 to 109

RACE=NW SEX=F